2013

Building and simulating protein machines

Ataur Rahim Katebi
Iowa State University

Building and simulating protein machines

by

Ataur Rahim Katebi

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Robert L. Jernigan, Co-Major Professor

Drena Dobbs, Co-Major Professor

Vasant Honavar

Amy Andreotti

Alicia Carriquiry

Iowa State University
Ames, Iowa
2013

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS

First I like to thank my major professor Dr. Robert L. Jernigan for his guidance and support throughout this process. His encouragement to think beyond the limit helped me to investigate into such an interdisciplinary problem in this new and fascinating area of research. I also like to thank my co-major professor Dr. Drena Dobbs and other POS committee members Drs. Vasant Honavar, Amy Andreotti, and Alicia Carriquiry, for their support, especially while I was at the beginning phase of this journey. Dr. Mark N. Gleason who has been my PFF mentor and Dr. Holly Bender, the PFF director, helped me to improve my teaching philosophy, scientific communication, and interpersonal communication through the Program for Future Faculty (PFF). I am immensely indebted to all the above.

My learning has been facilitated through the interaction with the present and past members of the Jernigan laboratory throughout these years – Kanan Sarkar, Liu Jie, Kejue Jia, Yuan Wang, Debkanta Chakrabarty, and Drs. Guang Song, Taner Z. Sen, Xuefeng Zhao, Michael T. Zimmermann, Scott Boykon, and Sumudu Leelananda.

Trish Stauble (BCB program), Mary Jane McCunn (LH Baker Center), Karen Bovenmyer (Center for Excellence in Learning and Teaching), and Andrea Dinkelman (Parks Library) have helped me to make the graduate studies at Iowa State University smoother.

Throughout this process my invaluable friends Ismail Siti Izera, Dr. Aftab Alam, Dr. Mohammad Al Hasan, Lailatul Ali Husin, Rukshana Ruby, and my sisters Nasima Katebi and Nusrat Katebi provided their kindness and support.

ABSTRACT

Glycolysis is a central metabolic pathway, present in almost all organisms, that produces energy. The pathway has been extensively investigated by biochemists. There is a significant body of structural and biochemical information about this pathway. The complete pathway is a ten step process. At each step, a specific chemical reaction is catalyzed by a specific enzyme. Fructose bisphosphate aldolase (FBA) and triosephosphate isomerase (TIM) catalyze the fourth and the fifth steps on the pathway.

This thesis investigates the possible substrate transfer mechanism between FBA and TIM. FBA cleaves its substrate, the six-carbon fructose-1,6-bisphosphate (FBP), into two three-carbon products – glyceraldehydes 3-phosphate (GAP) and dihydroxy acetone phosphate (DHAP). One component of these two products, DHAP, is the substrate for TIM and the other component GAP goes directly to GAPDH, the subsequent enzyme on the pathway. TIM converts DHAP to GAP and delivers the product to GAPDH. I employ Elastic Network Models (ENM) to investigate the mechanistic and dynamic aspects of the functionality of FBA and TIM enzymes – (1) the effects of the oligomerization of these two enzymes on their functional dynamics and the coordination of the individual protein's structural components along the functional region; and (2) the mechanistic synchrony of these two protein machines that may enable them to operate in a coordinated fashion as a conjugate machine – transferring the product from FBA as substrate to TIM.

A macromolecular machine comprised of FBA and TIM will facilitate the substrate catalysis mechanism and the product flow between FBA and TIM. Such a machine could

be used as a functional unit in building a larger a machine for the structural modeling of the whole glycolysis pathway. Building such machines for the glycolysis pathway may reveal the interplay of the enzymes as a complete machine. Also the methods and insights developed from the efforts to build such large machines could be applied to build macromolecular structures for other biologically important clusters of interacting enzymes centered around individual metabolic pathways.

CHAPTER 1. INTRODUCTION

1.1 Background, Problem Definition, and Significance

Glycolysis is a central metabolic pathway, which is present in almost all organisms, that supplies energy to the organism. The pathway has been extensively investigated by biochemists. There is a lot of structural and biochemical information about this pathway. The complete pathway is a ten step process. At each step, a specific chemical reaction is catalyzed by a very specific enzyme. Figure 1 shows a schematic diagram of this pathway – the arrow for a step is labeled by the enzyme that catalyzes the step; and the tail and the head of the arrows are labeled with the substrate and the product of the corresponding enzyme. The individual enzymes of the pathway are exceptionally well characterized – both in terms of their properties and their detailed structures. In this pathway, a six-carbon sugar glucose breaks down to create two molecules of pyruvate which is a three-carbon molecule. Along the way, 2 ATP molecules are used and 4 ATP molecules are generated. So the net energy produced in the glycolysis pathway is 2 units of ATP. The generated pyruvate can feed into three other subsequent metabolic cycles. In presence of oxygen, the pyruvate can be converted into acetyl-CoA which enters the Krebs cycle where acetyl-CoA is completely oxidized to generate more ATPs – in mammalian cells, 36 units of ATP per glucose is generated through oxidative phosphorylation. In the absence of oxygen, fermentation takes place, generating only 2 units of ATP per glucose molecule. In yeast, fermentation produces ethanol and carbon dioxide from pyruvate. In mammalian muscle, fermentation generates lactic acid which can result from strenuous exertion. This pathway is extremely important for an organism and failure in any one of its steps can have lethal effects on a cell.

The shaded region of Fig. 1 corresponds to the activities of FBA and TIM, the two enzymes that we are investigating in this research. The two steps in the shaded region, the fourth and the fifth steps, shows that FBA breaks down the six-carbon sugar FBP into two smaller three-carbon fragments – GAP and DHAP. The first of these is the proper substrate for the next step, but the DHAP is not, and so there is an additional enzyme recruited to convert it to the desired product GAP. This is a remarkable step to ensure the high efficiency of the process. The enzyme for this conversion step is TIM. The dynamics of this has been studied previously [2;3]. The structures of FBA and TIM exist in two essential forms – an open and a closed form, with a loop closing over the active site. These two enzymes must pair together for DHAP to transfer from FBA to TIM.

Several questions immediately arise. How do the opening and closing of these two enzymes facilitate the transfer of substrate? Do these two enzymes bind together for the substrate transfer? Is there a direct tunnel for transferring the substrate? Are the reactions of these two enzymes synchronized? How does functional mechanism of the machinery between these enzymes synchronize with the other two enzymes – the upstream PFK and the downstream GAPDH? In this research, we address in detail this structural interdependence for the activity of these enzymes. We also study the dynamics of these proteins to get a deeper understanding of how these sequentially active enzymes efficiently carry out their required work. We use computational methods to investigate the patterns of dynamics of FBA and TIM upon oligomerization across different species. We also investigate how the motions of these two proteins are correlated, especially for their functionally significant regions. From my research on the individual proteins and also about their coordination, I want to learn how these two proteins mechanistically coordinate their functions in a joint fashion to transfer product from FBA to TIM.

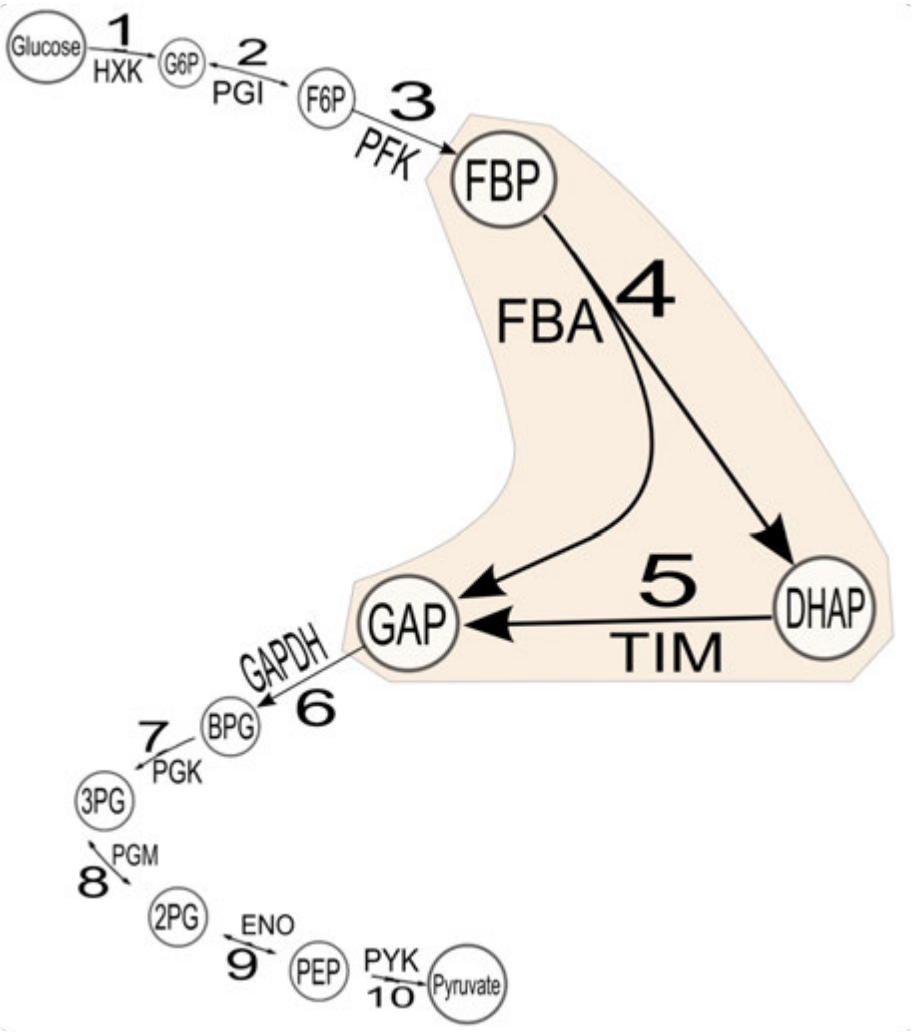


Figure 1. (A) Glycolysis pathway. Adapted from [1]. The abbreviations for the enzymes are as follows: HXK – hexokinase; PGI – glucose phosphate isomerase; PFK – phosphofructokinase; FBA – fructose 1,6-bisphosphate aldolase; TIM – triosephosphate isomerase; GAPDH – glyceraldehyde-phosphate dehydrogenase; PGK – phosphoglycerate kinase; PGM – phosphoglycerate mutase; ENO – enolase; PYK – pyruvate kinase. The abbreviations for the substrates and prodcuts (marked with circles) are as follows: G6P – glucose 6-phosphate; F6P – fructose 6-phosphate; FBP – fructose 1,6-bisphosphate; GAP – glyceraldehyde 3-phosphate; BPG – 1,6-bisphospho glycerate; 3PG – 3-phosphoglycerate; 2PG – 2-phosphoglycerate; PEP – phosphoenol pyruvate. The numbers on the arrows indicate the steps of the pathways.

The shaded region of Fig. 1 corresponds to the activities of FBA and TIM, the two enzymes that we are investigating in this research. The two steps in the shaded region, the fourth and the fifth steps, shows that FBA breaks down the six-carbon sugar FBP into two smaller three-carbon

fragments – GAP and DHAP. The first of these is the proper substrate for the next step, but the DHAP is not, and so there is an additional enzyme recruited to convert it to the desired product GAP. This is a remarkable step to ensure the high efficiency of the process. The enzyme for this conversion step is TIM. The dynamics of this has been studied previously [2;3].

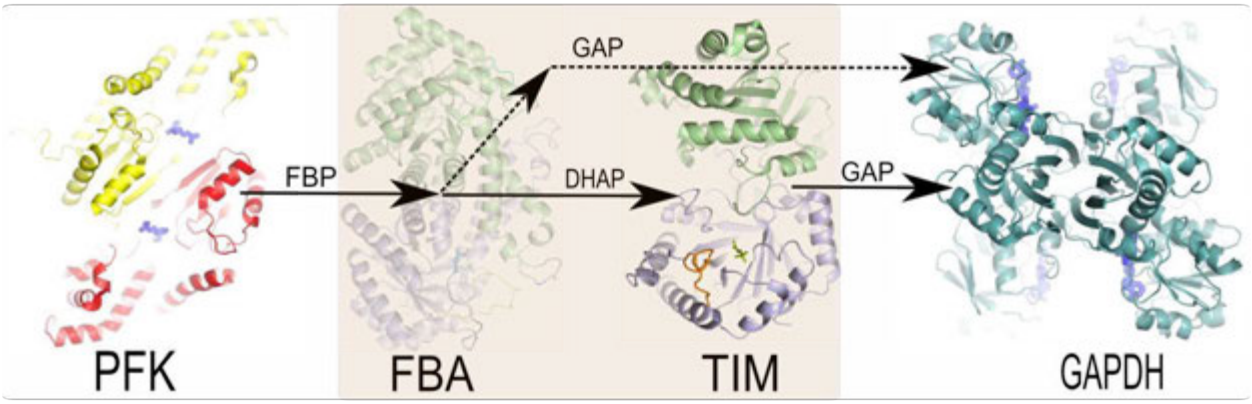


Figure 2. Flow of the substrates and the products from PFK to FBA, from FBA to TIM, and from FBA & TIM to GAPDH. Acronyms are defined in the caption for Fig. 1. The shaded section is the principal concern of this research.

This thesis investigates the substrate transfer mechanism between FBA and TIM. FBA cleaves its substrate, six-carbon FBP, into two three-carbon products – GAP and DHAP. One component of these two products, DHAP, becomes substrate to TIM and the other component GAP directly goes to GAPDH, the subsequent enzyme on the pathway. TIM converts DHAP to GAP and delivers the product to GAPDH. Figure 2 shows the schematic diagram of the flow of substrates and products from PFK to FBA, from FBA to TIM, and from FBA and TIM to GAPDH. In this research, I investigate the mechanistic and dynamic aspects of the functionality of FBA and TIM enzymes – (1) the effects of the oligomerization of these two enzymes on their functional dynamics and the coordination of the individual protein's structural components along the functional region; and (2) the mechanistic synchrony of these two protein machines that may enable them to operate in a coordinated fashion as a conjugate machine – transferring product of FBA as substrate to TIM.

A macromolecular machine comprised of FBA and TIM might facilitate the substrate catalysis mechanism and the product flow between FBA and TIM. Such a machine could be used as a functional unit in building larger a machine including the four enzymes as shown in Fig. 2, which would further be informative in the structural modeling of the whole glycolysis pathway. Building such machines for the glycolysis pathway may reveal the interplay of the enzymes as a complete machine. Also the methods and insights developed from the efforts to build such large machines could be applied to build macromolecular structures for other biologically important clusters of interacting proteins centered around individual metabolic pathways.

Interactions and coordinated movements between proteins are essential to carry out many different biological functions in an organism. A protein-protein interaction network consists of thousands of proteins in the organism and hundreds of thousands of interactions among them. All these proteins function as part of an intricate network of physical complexes and pathways [4]. Several databases have been developed to record the protein-protein interactions that have been found by employing different experimental methods. Some of these interaction databases are Biogrid [5], MIPS [6], BIND [7], DIP [8], MINT [9], etc. These networks include thousands of proteins and hundreds of thousands of interactions, such as Biogrid Network 2.0.41 that consists of 5,425 proteins and 121,664 interactions among them. Despite the presence of so many proteins and interactions in a PPIN, the proteins in a network form functional clusters of proteins where the proteins in a certain cluster perform some specific biological function in a concerted and coordinated manner such as proteins on the glycolysis pathway as shown in Fig. 1. There are different clustering methods for grouping the proteins in a PPIN; some of such clustering methods are – Molecular complex detection (MCODE) [10], Markov chain clustering (MCL)[11], clustering based on network distance [12] such as UVCLUSTER[13] , unsupervised

graph clustering [14;15] such as a two-step approach for clustering [16], Super paramagnetic clustering (SPC) [17;18], Restricted Neighborhood Search Clustering Algorithm (RNSC) [19;20], etc. MCODE is very successful method to detect densely connected regions in a PPIN. It weights each vertex by the density of the local neighborhood. First, it chooses a few vertices with large weights and isolates the dense regions according to some parameters. The MCL method simulates a random walk within a graph where each node of the graph represents a protein and the edge between two nodes is weighted with the sequence similarity of the proteins that are represented by the two nodes. It groups the highly similar proteins into clusters. The UVCLUSTER is a distance based clustering method. In this method, association between every pair of nodes is calculated. The association of a vertex with itself is highest and the association between two vertices with no connecting path is defined to be 0. This method groups closely related proteins into clusters. In an unsupervised two step method, a suitable model is defined. First, the model is trained with the training set of the data, and second, the trained model is tested on the test set of the data. The SPC method is a temperature based method that finds tightly-connected nodes in a graph. The RNSC method partitions the graph based on a cost function which is defined as a function of invalid connections. An invalid connection incident with a node v is a connection that exists between v and a node in a different cluster. Biological pathways in a cell can be considered as some real examples of functional clusters of proteins.

On the other hand, the Protein Data Bank (PDB) [7;21] is developed as a depository for the structures of proteins and protein complexes that are determined by different experimental methods such as X-ray crystallography, NMR, electron microscopy, etc. the PDB had 81,155 structures as of 22 January 2013. There are also *ab initio* and template-based methods to predict protein structures. Some successful structure prediction tools are I-TASSER [22;23],

MODELLER [24], Rosetta [25;26], WHAT IF [27], and Phyre [28]. There are two kinds of structural modeling approaches – homology-based modeling where a template structure from the PDB is used. A template is a PDB structure that has high sequence similarity with the target sequence. Then this template is used as a guide to fold the target sequence into a 3D model. When a good template is not found, an *ab initio* modeling is used to fold the target sequence. *Ab initio* methods are based on prediction of a structure that may rely on a detailed energy function that defines the relationship between residues in the sequence. This method attempts to minimize this function to reach to lowest energy level of the structure by generating a range of structures. Moreover, there are computational methods that are developed to predict the complexes between proteins. Several such methods that rank well in the competition for complex prediction [29] are ClusPro [30], Z-Dock [31], Rosetta Dock [32], HADDOCK [33], FireDock [34], etc. In different approaches to the modeling of complexes, a scoring function is defined to model the interaction between residues of the two proteins. Then a huge number of docking poses are generated based on the scoring function. Then these poses are clustered and ranked. The best model may be found at the top ranked clusters. The Critical Assessment of Structure Prediction (CASP) experiment assesses the improvement of structural modeling every two years [35-38]. A similar experiment, The Critical Assessment of Predicted Interactions (CAPRI), assesses the improvement of modeling protein-protein interactions every two years [39]. Though these methods are still in the developmental phase, they can sometimes give good structural/docking models, and the community-wide effort to improve these methods continues.

The field of modeling protein dynamics has also made significant progress. Elastic Network Models (ENMs) can usually successfully capture the global motions of proteins [40-42]. In this method, a protein molecule is represented as a network of nodes where each node, denoting a

residue or a set of residue, is a mass and an edge between two nodes is denoted by a spring. Normal modes of this mass-spring network represent the motions of the macromolecular structure. These models can be developed for either atomic or coarse-grained models. The coarse-grained models are useful because dominant motions are normally the slowest motions, and these depend less on the atomic details and more on the global properties such as overall shape. Because of the significant improvements in computational power recently, all atom molecular dynamics programs such as GROMACS [43], CHARMM [44], AMBER [45], etc. can model the dynamics of proteins to study the finer atomic level motions of the protein structures. In these methods, the atoms and molecules are allowed to interact based on Newton's equations of motion for a system for a set of interacting particles. The forces and potential energy between the interacting atoms are defined by molecular mechanics force fields. Improved computational power and enhanced error handling algorithms for the integration of force fields over longer time simulations, make it possible to observe the fine-grained dynamics of large macromolecular systems.

Considering these recent developments in the above areas, we are approaching a period in computational biological science, where building machines for clusters of proteins for parts of the protein-protein interaction network (PPIN) is a next natural scientific step. My research in this dissertation gives a framework for building such machines across PPIN.

1.2 Dissertation Organization

Figure 3 shows the organization of this dissertation. Chapters 2 and 3 provide the detailed analysis of change of motions over oligomerization of FBA and TIM proteins, respectively. Chapter 4 discusses the issues to build models for the complex between FBA and TIM. Chapter 5 provides generic methods to cluster a protein-protein interaction network (PPIN) and build models for the complexes in the clusters. Chapter 6 summarizes the results from chapters 2, 3 and 4 in a holistic way, discusses the application of the results in solving the proposed research problem, and proposes future research directions.

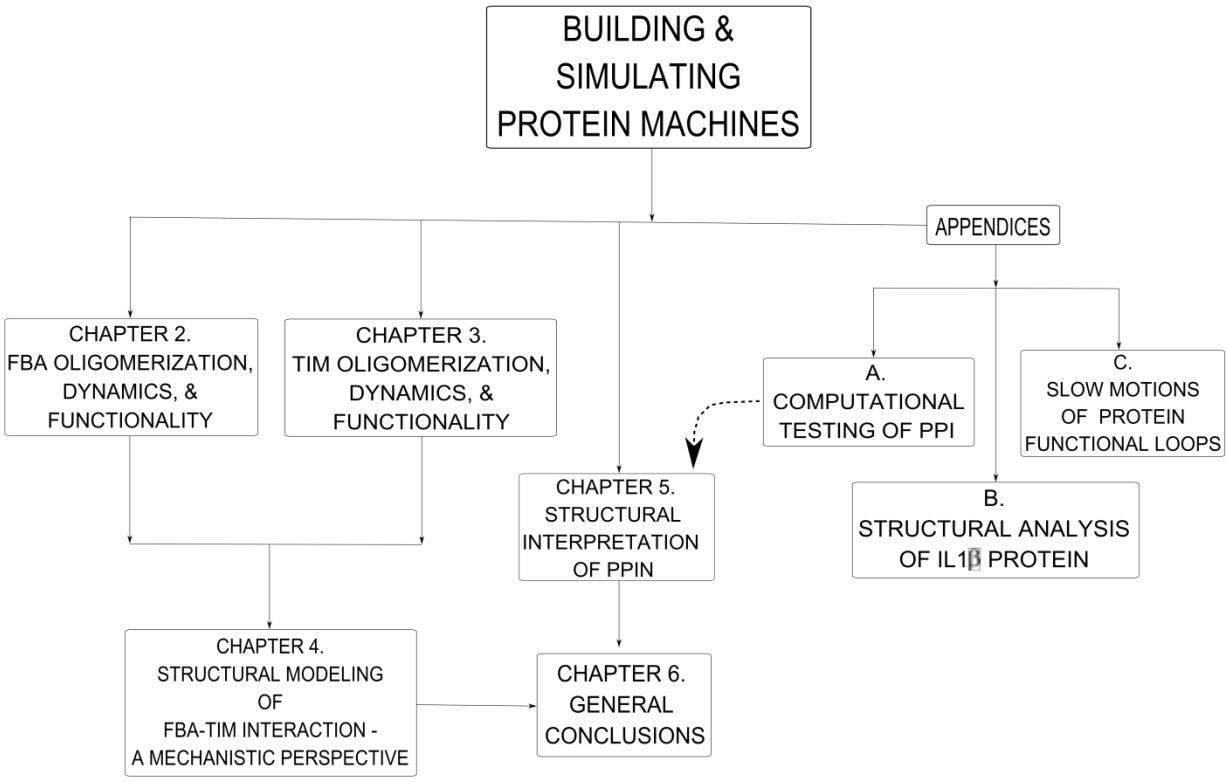


Figure 3. Dissertation Organization.

Chapter 2: Aldolase Oligomerization Relates to Specific Dynamics Essential to Carry out its Function

This dissertation chapter is a manuscript prepared for submission to a peer reviewed journal. This chapter describes the details of the available structures and their oligomerization sites. It goes ahead to analyze how the dynamics of fructose -1,6-bisphosphate (FBA) and tagatose-1,6 bisphosphate (TBA) proteins, two class II aldolases, change upon their oligomerization. FBA condenses dihydroxy acetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP) into fructose bisphosphate (FBP) and vice versa. TBA condenses dihydroxy acetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP) into tagatose bisphosphate (TBP) and vice versa. FBA and TBA are two similar proteins on the glycolysis and galactose metabolism pathways. We focus on three structures – one dimeric FBA structure and one tetrameric TBA structure from the mesophilic *E.coli*, and one tetrameric FBA structure from the hyperthermophilic *T.aquaticus* organism. In this part of research, we find that oligomerization of aldolases helps the structures to stabilize the oligomerization interfaces and achieve the important consistent functional dynamics across the catalytic regions.

Chapter 3: Triosephosphate isomerase Structure Space Diversity: Oligomerization, Dynamics, and Functionality – An Evolutionary Perspective

This chapter is also a manuscript prepared for submission to a peer reviewed journal. It presents a description and analysis of triosephosphate isomerase (TIM) structures – the fifth enzyme on the glycolysis pathway. Mesophilic TIM structures are dimeric but hyperthermophilic structures are tetrameric. The basic research question is: how does the TIM oligomerization affect the functional dynamics of this protein and its stability? To answer that

question, I investigate into four TIM structures – one engineered monomeric TIM (monoTIM), a dimeric TIM from mesophilic *T.brucei* (TbTIM), two tetrameric TIM structures (TmTIM and PwTIM) from *T.maritima* and *P.woesei*, respectively. We find that oligomerization not only stabilizes the structures, it also increases their functional dynamics. Moreover, the functional loops are highly coordinated with each other such that it helps the opening and closing of the catalytic pocket.

Thus the stabilized dimeric and tetrameric TIM structures achieve high turn-out rates through their highly coordinated dynamics of the functional loops.

Chapter 4: Structural Modeling of Fructose bisphosphate aldolase and Triosephosphate isomerase Interaction – A Mechanistic Perspective

This chapter is a manuscript prepared for submission to a peer reviewed journal as well. This chapter brings together the findings in chapters 2 and 3. First, we compare the structural features of FBA and TIM, the two enzymes investigated in chapters 2 and 3, respectively, and find that the cores of these two enzymes are highly similar (RMSD 4.8 Å). The functional loops also align properly. By comparing the dynamics of the functional loops within and between these enzymes, we find that the dynamics of these loops are well coordinated within and between the two structures. Both enzymes have 'phosphate gripper' region on one of the functional loops – the tip of loop 6 in the TIM structure and the N-terminus of loop 6 in the FBA structure. FBA could use its 'phosphate gripper' to hunt its substrate from the surroundings or the product of the previous enzyme on the pathway and similarly TIM could use its 'phosphate gripper' to import the FBA product into its catalytic pocket. High synchrony of the motions of the functional loops within the structures and between the structures, and the 'phosphate gripper' motif on their functional loop 6 indicates that these two enzymes could form a conjugated FBA-TIM machine.

Chapter 5: Structural Interpretation of Protein-Protein Interaction Network

This chapter has been published in the peer reviewed scientific journal, BMC Structural Biology. This chapter deals with the concept of building models for the complexes across the protein-protein interactome. This chapter has two main components. First, we use computational methods to cluster the protein-protein interaction network, where each cluster is a set of proteins that are functionally relevant. Second, we attempt to build models for the complexes of those proteins within a cluster.

Appendix A. Computational Testing of Protein-protein Interactions

This appendix is a published manuscript in the peer reviewed conference proceedings, *IEEE BIBM*. In this paper, we cluster yeast protein-protein interaction network and investigate the interactions in the individual clusters for relationships among the member proteins. We also build the 3D structural models of the proteins whose structures are not available in the protein data bank. We explore ways to predict unreported interactions by docking the protein structures in the clusters.

Appendix B. Immunological Implication of Structural Analysis of Porcine IL1β Proteins Expressed in Macrophages and Embryos

This appendix is a published manuscript in the peer reviewed conference proceedings, *ACM* International *Conference on Bioinformatics and Computational Biology*. IL1β is an important vertebrate animal protein. It is a member of the cytokine protein family and is involved in generating an inflammatory response to some infections. Researchers have found that there are

two porcine IL1β proteins expressed – one in embryos and a different one in macrophage and endometrial tissues. However, these two proteins have about 86% sequence identity. In this paper, we describe how these two proteins might be structurally and functionally different. We find interesting aspects of these two structures that differ: 1) A predicted binding site appears to have different side chain arrangements that might lead to different binding efficiencies for the same protein or even to different partners. 2) The Caspase 1 cleavage site in the precursor proteins differs in a way that has previously been experimentally determined to be important and to reduce the cleavage activity by one order of magnitude for the embryonic IL1β, conferring a significant advantage to the protein (embryonic IL1β).

Appendix C. The Importance of Slow Motions for Protein Functional Loops

This appendix is a published manuscript in the peer reviewed journal, Physical Biology. Loops in proteins connect secondary structures such as alpha-helix and beta-sheet, are often on the surface, and may play a critical role in some functions of a protein. The mobility of loops is central for the motional freedom and flexibility requirements of active-site loops and may play a critical role for some functions. The structures and behaviors of loops have not been much studied in the context of the whole structure and its overall motions, and especially how these might be coupled. The loop motions are investigated by using coarse-grained structures (C^{α} atoms only) to solve for the motions of the system by applying Lagrange equations with elastic network models to learn about which loops move in an independent fashion and which move in coordination with domain motions, faster and slower, respectively. The normal modes of the system are calculated using eigen-decomposition of the stiffness matrix. The contribution of individual modes and groups of modes are investigated for their effects on all residues in each loop by using Fourier analyses. Our results indicate overall that the motions of functional sets of

loops behave in similar ways as the whole structure. But, overall only relatively few loops move in coordination with the dominant slow modes of motion, and these are often closely related to function.

Bibliography

[1] L. A. Fothergill-Gilmore and P. A. Michels, "Evolution of glycolysis," *Prog. Biophys. Mol. Biol.*, vol. 59, no. 2, pp. 105-235, 1993.

[2] Kurkcuoglu O, Jernigan RL, and Pemra D, "Collective Dynamics of Large Proteins from Mixed Coarse-Grained Elastic Network Model," *QSAR & Combinatorial Science*, vol. 24, pp. 443-448, Jun. 2005.

[3] O. Kurkcuoglu, R. L. Jernigan, and P. Doruker, "Loop motions of triosephosphate isomerase observed with elastic networks," *Biochemistry*, vol. 45, no. 4, pp. 1173-1182, Jan. 2006.

[4] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, "Evolution of the protein repertoire," *Science*, vol. 300, no. 5626, pp. 1701-1703, Jun. 2003.

[5] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. Database issue, p. D535-D539, Jan. 2006.

[6] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H. W. Mewes, A. Ruepp, and D. Frishman, "The MIPS mammalian protein-protein interaction database," *Bioinformatics.*, vol. 21, no. 6, pp. 832-834, Mar. 2005.

[7] G. D. Bader, D. Betel, and C. W. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 248-250, Jan. 2003.

[8] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, no. Database issue, p. D449-D451, Jan. 2004.

[9] A. Ceol, A. A. Chatr, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni, "MINT, the molecular interaction database: 2009 update," *Nucleic Acids Res.*, vol. 38, no. Database issue, p. D532-D539, Jan. 2010.

[10] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC. Bioinformatics.*, vol. 4, p. 2, Jan. 2003.

[11] Stijn van Dongen, "A stochastic uncoupling process for graphs.," 2000.

[12] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Proc. Natl. Acad. Sci U. S. A*, vol. 100, no. 3, pp. 1128-1133, Feb. 2003.

[13] V. Arnau, S. Mars, and I. Marin, "Iterative cluster analysis of protein interaction data," *Bioinformatics.*, vol. 21, no. 3, pp. 364-378, Feb. 2005.

[14] C. A. Ouzounis, R. M. Coulson, A. J. Enright, V. Kunin, and J. B. Pereira-Leal, "Classification schemes for protein structure and function," *Nat. Rev. Genet.*, vol. 4, no. 7, pp. 508-519, Jul. 2003.

[15] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis, "Detection of functional modules from protein interaction networks," *Proteins*, vol. 54, no. 1, pp. 49-57, Jan. 2004.

[16] P. Pei and A. Zhang, "A Two-Step Approach for Clustering Proteins based on Protein Interaction Profile," *Proc. IEEE Comput. Syst. Bioinform. Conf.*, vol. 2005, no. 1544467, pp. 201-209, 2005.

[17] M. Blatt, S. Wiseman, and E. Domany, "Superparamagnetic clustering of data," *Phys. Rev. Lett.*, vol. 76, no. 18, pp. 3251-3254, Apr. 1996.

[18] G. Getz, E. Levine, and E. Domany, "Coupled two-way clustering analysis of gene microarray data," *Proc. Natl. Acad. Sci U. S. A*, vol. 97, no. 22, pp. 12079-12084, Oct. 2000.

[19] A. D. King, N. Przulj, and I. Jurisica, "Protein complex prediction with RNSC," *Methods Mol. Biol.*, vol. 804, pp. 297-312, 2012.

[20] A. D. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics.*, vol. 20, no. 17, pp. 3013-3020, Nov. 2004.

[21] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, Jan. 2000.

[22] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC. Bioinformatics.*, vol. 9, p. 40, 2008.

[23] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nat. Protoc.*, vol. 5, no. 4, pp. 725-738, Apr. 2010.

[24] N. Eswar, D. Eramian, B. Webb, M. Y. Shen, and A. Sali, "Protein structure modeling with MODELLER," *Methods Mol. Biol.*, vol. 426, pp. 145-159, 2008.

[25] R. Das and D. Baker, "Macromolecular modeling with rosetta," *Annu. Rev. Biochem.*, vol. 77, pp. 363-382, 2008.

[26] R. Das, I. Andre, Y. Shen, Y. Wu, A. Lemak, S. Bansal, C. H. Arrowsmith, T. Szyperski, and D. Baker, "Simultaneous prediction of protein folding and docking at high resolution," *Proc. Natl. Acad. Sci. U. S. A*, vol. 106, no. 45, pp. 18978-18983, Nov. 2009.

[27] G. Vriend, "WHAT IF: a molecular modeling and drug design program," *J. Mol. Graph.*, vol. 8, no. 1, pp. 52-6, 29, Mar. 1990.

[28] L. A. Kelley and M. J. Sternberg, "Protein structure prediction on the Web: a case study using the Phyre server," *Nat. Protoc.*, vol. 4, no. 3, pp. 363-371, 2009.

[29] J. Janin, K. Henrick, J. Moult, L. T. Eyck, M. J. Sternberg, S. Vajda, I. Vakser, and S. J. Wodak, "CAPRI: a Critical Assessment of PRedicted Interactions," *Proteins*, vol. 52, no. 1, pp. 2-9, Jul. 2003.

[30] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics.*, vol. 20, no. 1, pp. 45-50, Jan. 2004.

[31] R. Chen and Z. Weng, "Docking unbound proteins using shape complementarity, desolvation, and electrostatics," *Proteins*, vol. 47, no. 3, pp. 281-294, May 2002.

[32] S. Lyskov and J. J. Gray, "The RosettaDock server for local protein-protein docking," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, p. W233-W238, Jul. 2008.

[33] C. Dominguez, R. Boelens, and A. M. Bonvin, "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information," *J Am. Chem Soc.*, vol. 125, no. 7, pp. 1731-1737, Feb. 2003.

[34] N. Andrusier, R. Nussinov, and H. J. Wolfson, "FireDock: fast interaction refinement in molecular docking," *Proteins*, vol. 69, no. 1, pp. 139-159, Oct. 2007.

[35] V. Mariani, F. Kiefer, T. Schmidt, J. Haas, and T. Schwede, "Assessment of template based protein structure predictions in CASP9," *Proteins*, vol. 79 Suppl 10, pp. 37-58, 2011.

[36] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP)--round IX," *Proteins*, vol. 79 Suppl 10, pp. 1-5, 2011.

[37] A. Kryshtafovych, K. Fidelis, and J. Moult, "CASP9 results compared to those of previous CASP experiments," *Proteins*, vol. 79 Suppl 10, pp. 196-207, 2011.

[38] A. Kryshtafovych, C. Venclovas, K. Fidelis, and J. Moult, "Progress over the first decade of CASP experiments," *Proteins*, vol. 61 Suppl 7, pp. 225-236, 2005.

[39] J. Janin, "Protein-protein docking tested in blind predictions: the CAPRI experiment," *Mol. Biosyst.*, vol. 6, no. 12, pp. 2351-2362, Dec. 2010.

[40] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophys. J.*, vol. 80, no. 1, pp. 505-515, Jan. 2001.

[41] I. Bahar, A. R. Atilgan, and B. Erman, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential," *Fold. Des*, vol. 2, no. 3, pp. 173-181, 1997.

[42] I. Bahar, B. Erman, T. Haliloglu, and R. L. Jernigan, "Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations," *Biochemistry*, vol. 36, no. 44, pp. 13512-13523, Nov. 1997.

[43] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics.*, Feb. 2013.

[44] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: the biomolecular simulation program," *J Comput. Chem*, vol. 30, no. 10, pp. 1545-1614, Jul. 2009.

[45] E. Darian and P. M. Gannett, "Application of molecular dynamics simulations to spin-labeled oligonucleotides," *J Biomol. Struct. Dyn.*, vol. 22, no. 5, pp. 579-593, Apr. 2005.

CHAPTER 2. ALDOLASE OLIGOMERIZATION RELATES TO SPECIFIC DYNAMICS ESSENTIAL TO CARRY OUT ITS FUNCTION

Manuscript prepared for submission to a peer reviewed scientific journal

Ataur R Katebi and Robert L Jernigan

LH Baker Center

Iowa State University

Abstract

Aldolases are enzymes that catalyze the reversible reaction to cleave fructose/tagatose 1,6-bisphosphate into two triose phosphate components – dihydroxy acetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP). There are Schiff base (class I) and metal base (class II) aldolases. Class II aldolases that can be dimeric or tetrameric structures. Dimerization occurs through a type I interface and two homo dimeric structures join together to form a tetrameric aldolase structure by interactions across two type II interfaces. We apply ENM to investigate the dynamics of three aldolase structures – *E.coli* fructose 1,6-bisphosphate aldolase (FBA), *E.coli* tagatose 1,6-bisphosphate aldolase (TBA), and *T.aquaticus* FBA. We find that oligomerization not only stabilizes the aldolase structures along the interface region, but also helps the protein to achieve required dynamics for the functional loops. The acquired mobility of the functional loops facilitates the mobility of the residues that constitute the catalytic microenvironment that is important for catalysis. to sample the important.

Key Words: fructose 1,6-bisphosphate aldolase; tagatose 1,6-bisphosphate aldolase; fructose 1,6-bisphosphate; tagatose 1,6-bisphosphate; dihydroxy acetone phosphate; glyceraldehyde 3-phosphate; triosephosphate isomerase.

Abbreviations:

PGH	phosphoglycolohydroxamic acid	TBP	tagatose-1,6 bisphosphate
DHAP	dihydroxy acetone phosphate	TIM	triosephosphate isomerase
GAP	glyceraldehyde 3-phosphate	PDB	Protein Data Bank
FBA	fructose 1,6-bisphosphate aldolase	ENM	Elastic Network Model
FBP	fructose 1,6-bisphosphate	ANM	Anisotropic Network Model
TBA	tagatose 1,6-bisphosphate aldolase		

2.1 Introduction

2.1.1 Class I and Class II FBAs

FBA is the fourth enzyme in the glycolysis pathway which is considered to be one of the earliest pathways. This enzyme catalyzes the reversible condensation of two three carbon sugar substrates – DHAP and GAP into a six carbon fructose 1,6-bisphosphate (FBP). There are two mechanistically distinct types of aldolases: Class I and Class II [1]. Class I aldolases form a covalent Schiff-base intermediate between DHAP moiety of the substrate and an ε-amino group of an active site lysine residue during the catalysis [2;3]. On the other hand, class II aldolases employ a divalent cation – usually Zn^{+2} or Fe^{+2} as an electrophile in the catalytic cycle [2].

2.1.2 Diversity and Conservation in FBA Sequence and Structure Space

FBA has a relatively low sequence conservation across different species. However, the key residues in both class I and class II structures are conserved. Figure 1 shows a multiple sequence alignment of class II FBA enzymes from nine organisms – *Bacillus anthracis* (*B.anthracis*), *Campylobacter jejune* (*C.jejune*), *Coccidioides immitis* (*C.immitis*), *Escherichia coli* (*E.coli*),

Giardia lamblia (*G.lamblia*), *Mycobacterium tuberculosis* (*M.tuberculosis*), *Seccharomyces cerevisiae* (*S.cerevisiae*), *Thermus aquaticus (T.aquaticus*), and *Thermus caldophilus* (*T.caldophilus*), which reveals the conserved amino acids in class II enzymes. Table I shows the wide variety of scores from MSA – as small as 18.0 (between *E.coli* and *G.lamblia*) and as large as 92.0 (between *T.aquaticus* and *T.caldophilus*).

Figure 2 shows how the pair wise RMSD values of the subunit structures change according to the pairwise sequence alignment scores for the nine class II FBAs. It is evident that alignment scores and RMSD values are not perfectly anticorrelated. For some pairs, as sequence alignment scores increase, the structure alignment scores (RMSD values) do not decrease. However, the lowest RMSD (1.01 Å) corresponds to the pair with the highest sequence alignment score (92.0) – pair (8:9 – *T.aquaticus*: *T.caldophilus*). The largest RMSD score (5.77 Å) is observed for the pair (1:3 – *B.anthracis: C.jejune*) whose sequence alignment score is 26.0 – not the smallest score which is 18.0 for the pair (4:5 – *E.coli: G.lamblia*). This says that overall the sequence alignment scores are anti-correlated with the structural alignment scores but that this correlation is not strictly linear.

Moreover, although *T.aquaticus* and *T.caldophilus* have a high sequence alignment score (92.0) and a low RMSD score (1.01 Å), *T.caldophilus* achieves a different kind of functionality – it can synthesize both FBA and TBP from the same substrate components – DHAP and GAP. This indicates that even though it is in the same class of enzymes, a small difference in the sequence can lead to different functionality.

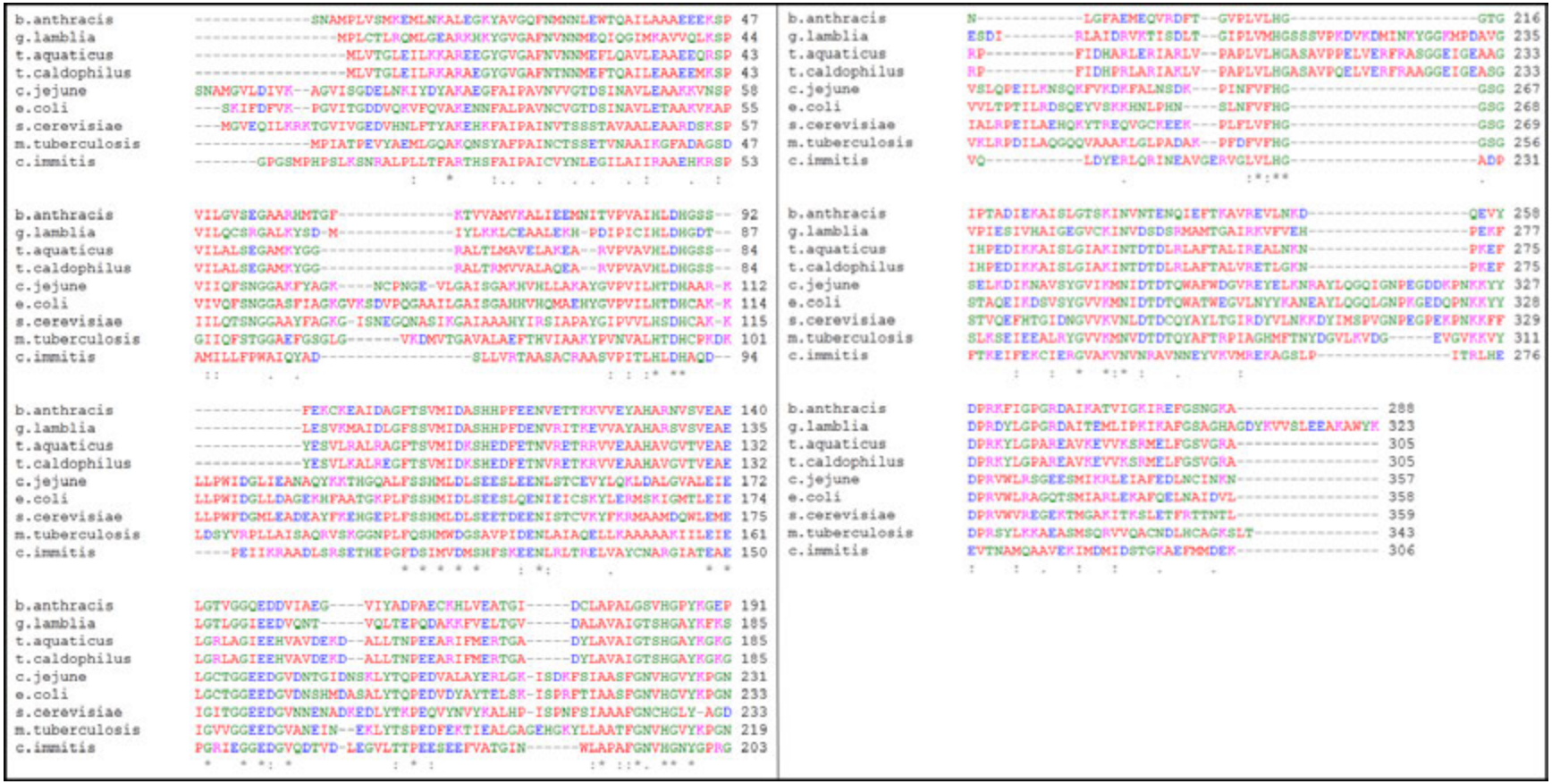


Figure 1. Multiple sequence alignment of nine class II fructose 1,6-bishphosphate aldolase sequences by using clustalW [4]. Meaning of symbols and colors are described below. Consensus symbols: * (asterisk): positions with a single, fully conserved residue; :(colon) – conservation between amino acids having closely similar characteristics – scoring > 0.5 in the Gonnet PAM 250 matrix [5]; .(period) – conservation between amino acids with weak similar characteristics– scoring =< 0.5 in the Gonnet PAM 250 matrix.
Meaning of the colored amino acids: RED small and hydrophobic; BLUE Acidic; MAGENTA Basic ; GREEN Hydroxyl + sulfhydryl + amine; Grey Unusual amino/imino acids etc.

Table I. Pair-wise Sequence Alignment Scores (SAC) and Subunit RMSD Values for nine Class II fructose 1,6-bishphosphate aldolases (Sorted by SAC)					
(id 1:id 2) (organism 1:organism 2)	Alignment Scores		(id 1:id 2) (organism 1:organism 2)	Alignment Scores	
	Sequence	Structure (Å)		Sequence	Structure (Å)
(4:5 - *E.coli: G.lamblia*)	18	2.75	(1:6 - *B.anthracis: M.tuberculosis*)	28	2.93
(2:4 - *C.immitis: E.coli*)	20	2.95	(3:8 - *C.jejune: T.aquaticus*)	28	4.31
(5:7 - *G.lamblia: S.cerevisiae*)	21	2.85	(3:9 - *C.jejune: T.caldophilus*)	28	4.11
(3:5 - *C.jejune: G.lamblia*)	22	3.83	(1:2 - *B.anthracis: C.immitis*)	29	2.58
(4:8 - *E.coli: T.aquaticus*)	22	3.98	(2:9 - *C.immitis: T.caldophilus*)	29	4.72
(4:9 - *E.coli: T.caldophilus*)	22	4.23	(2:8 - *C.immitis: T.aquaticus*)	30	3.76
(7:9 - *S.cerevisiae: T.caldophilus*)	22	3.89	(6:7 - *M.tuberculosis: S.cerevisiae*)	36	4.82
(2:3 - *C.immitis: C.jejune*)	23	4.15	(4:6 - *E.coli: M.tuberculosis*)	40	3.04
(7:8 - *S.cerevisiae: T.aquaticus*)	23	5.13	(3:6 - *C.jejune: M.tuberculosis*)	41	3.26
(1:4 - *B.anthracis: E.coli*)	24	2.97	(5:9 - *G.lamblia: T.caldophilus*)	46	1.52
(2:6 - *C.immitis: M.Tuberculosis*)	24	3.65	(1:8 - *B.anthracis: T.aquaticus*)	47	3.27
(5:6 - *G.lamblia: M.Tuberculosis*)	24	2.5	(3:7 - *C.jejune: S.cerevisiae*)	47	3.32
(2:7 - *C.immitis: S.cerevisiae*)	25	4.13	(4:7 - *E.coli: S.cerevisiae*)	47	2.18
(1:3 - *B.anthracis: C.jejune*)	26	5.77	(5:8 - *G.lamblia: T.aquaticus*)	47	1.53
(1:7 - *B.anthracis: S.cerevisiae*)	27	3.56	(1:5 - *B.anthracis: G.lamblia*)	48	1.59
(2:5 - *C.immitis: G.lamblia*)	27	3.3	(1:9 - *B.anthracis: T.caldophilus*)	48	3.18
(6:8 - *M.tuberculosis: T.aquaticus*)	27	4.36	(3:4 - *C.jejune: E.coli)*	63	3.66
(6:9 - *M.tuberculosis: T.caldophilus*)	27	3.56	(8:9 - *T.aquaticus: T.caldophilus*)	92	1.01

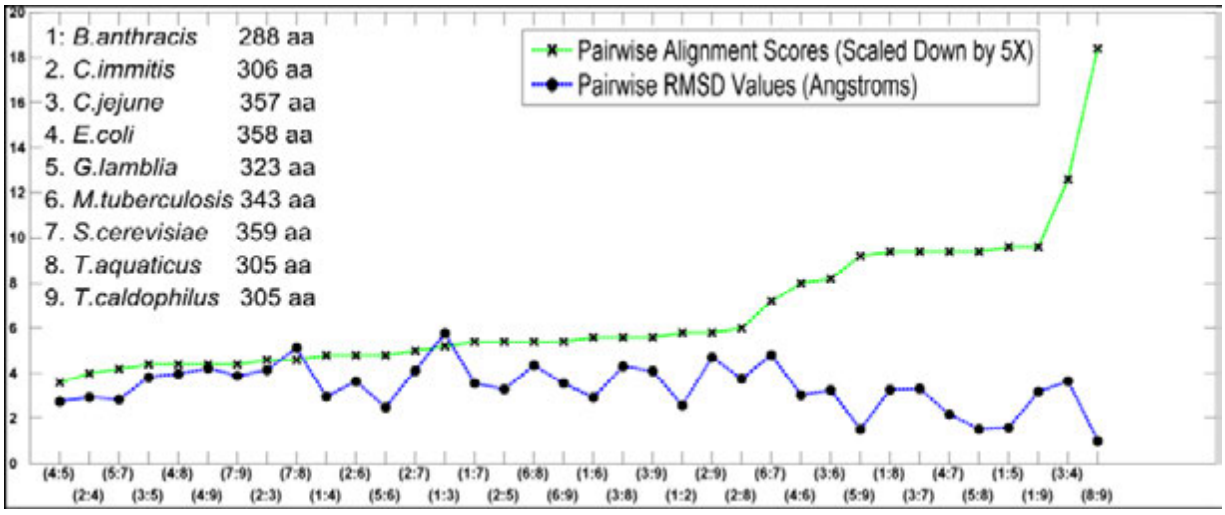


Figure 2. Comparison of sequence alignment and structure alignment results for Class II fructose 1,6-bishphosphate aldolases. Pair-wise alignment scores range from 18 to 92. RMSD values from 2.75 Å to 1.01 Å. Values are taken from Table I. While there is an overall trend for high sequence similarity to correlate with low RMSD values there are nonetheless significant deviations from such a simple relationship.

For the present work, we have selected three representative aldolase structures – *E.coli* FBA dimer, *E.coli* TBA tetramer, and *T.aquaticus* FBA tetramer. Table II shows the secondary structure placements in the three aldolases used to study the dynamics of the functional loops here. It is notable that the lengths of helix 2 and helix 3 are significantly shorter in *E.coli* TBA and *T.aquaticus* FBA than in *E.coli* FBA. Moreover, the segment of coil between helix 8A and helix 8B is also much longer in *E.coli* FBA than in the other two structures.

Panel A of Fig. 3 shows the architecture of a class II FBA subunit. It has an (α/β) barrel construction. Eight β-strands form the central barrel surrounded by helices α1 ~ α7, α8A, and α8B. Helix α8A and helix α8B stick out from the helical ring. The C-terminuses of the barrel are considered to be the front of the structure and the N-terminuses as the back side. There are loops on the front and back of the structure that connect the strands and the helices. The front loops (front loop 1 ~ front loop 8) connect sequentially from the strands to the helices – such as front loop 1 connnects β strand 1 to α helix 1, front loop 2 goes from β strand 2 to α helix 2, etc. The back loops (back loop 1 ~ back loop 7) connect sequentially from the helices to the strands – for example, back loop 1 connects from α helix 1 to β strand 2, back loop 2 from α helix 2 to β strand 2, and so on. Helix 0 (α0) is not one of the helices surrounding the central barrel; rather it covers the N-terminal opening at the back of the central β-barrel. There is a coil region between helices α8A and α8B. Panel B of Fig. 3 shows how the two subunits join along the interface region to construct the dimeric structure of the class II FBAs. Construction of this interface is named a type I interface. Panels C and D of Fig. 3 show two different views of a tetrameric class II FBA structure. In this configuration, two type I dimers bind together at two type II interfaces to form the tetrameric structure of a class II FBA.

Table II. Positions of the Secondary Structure Segments in the Sequences of *E.coli* FBA, *E.coli* TBA, and *T.aquaticus* FBA						
Type of Secondary Structure	**Sequence Indices in the Protein Sequence**					
	***E.coli* FBA**		***E.coli* TBA**		***T.aquaticus* FBA**	
	Indices	**Length**	**Indices**	**Length**	**Indices**	**Length**
N-terminal coil	1 – 15	15	1 – 6	6	1 – 5	5
N-terminal Helix 0	16 – 26	11	7 – 16	10	6 – 14	9
N-terminal Loop 0	27 – 30	4	17 – 19	3	15 – 18	4
α Helices						
Helix 1	40 – 52	14	29 – 41	13	28 – 39	12
Helix 2	80 – 100	20	60 – 71	12	58 – 70	13
Helix 3	116 – 133	18	88 – 94	7	86 – 92	7
Helix 4	151 – 165	15	111 – 125	15	109 – 123	15
Helix 5	199 – 209	11	156 – 166	11	156 – 164	9
Helix 6	239 – 252	14	191 – 200	10	191 – 200	10
Helix 7	272 – 278	7	216 – 222	7	236 – 244	9
Helix 8A	291 – 305	15	234 – 249	16	254 – 270	17
Coli between helix 8A & helix 8B	306 – 330	25	250 – 256	7	271 – 276	6
Helix 8B	331 – 352	22	257 – 278	22	277 – 299	23
Back Loops						
Loop 1	53 – 55	3	42 – 45	4	40 – 43	4
Loop 2	100 – 103	4	72 – 77	6	71 – 74	4
Loop 3	134 – 139	6	95 – 99	5	93 – 97	5
Loop 4	166 – 169	4	126 – 129	4	124 – 127	4
Loop 5	210 – 215	6	167 – 170	4	165 – 168	4
Loop 6	253 – 259	7	201 – 204	4	201 – 204	4
Loop 7	279 – 283	5	223 – 225	3	245 – 247	3
β Strands						
Strand 1	31 – 35	5	20 – 24	5	19 – 23	5
Strand 2	56 – 60	5	46 – 49	4	44 – 48	5
Strand 3	104 – 108	6	78 – 84	7	75 – 82	8
Strand 4	140 – 143	4	98 – 101	4	98 – 101	4
Strand 5	170 – 175	6	130 – 135	6	128 – 133	6
Strand 6	216 – 220	5	171 – 174	4	169 – 172	4
Strand 7	260 – 263	4	205 – 207	3	205 – 207	3
Strand 8	284 – 287	4	226 – 231	6	248 – 251	4
Front Loops						
Loop 1	35 – 39	3	25 – 28	4	24 – 27	4
Loop 2	61 – 79	20	50 – 59	10	49 – 57	9
Loop 3	109 – 115	7	85 – 87	3	83 – 85	3
Loop 4	144 – 150	7	102 – 110	9	102 – 108	7
Loop 5	176 – 198	23	136 – 155	20	134 – 155	22
Loop 6	221 – 238	18	175 – 190	26	173 – 190	18
Loop 7	264 – 271	8	208 – 215	8	208 – 235	28
Loop 8	288 – 290	3	232 – 233	2	252 – 253	2
C-terminal coil	353 – 358	6	279 – 285	7	300 – 305	6
Note: Segements with significant differences in lengths – helix 3, helix 6, coil between helix 8A and helix 8B, front loop 2, front loop 3, front loop 6, and front loop 7						

2.1.3 Monomer, Dimer, and Tetramer Architectures of Class II FBA

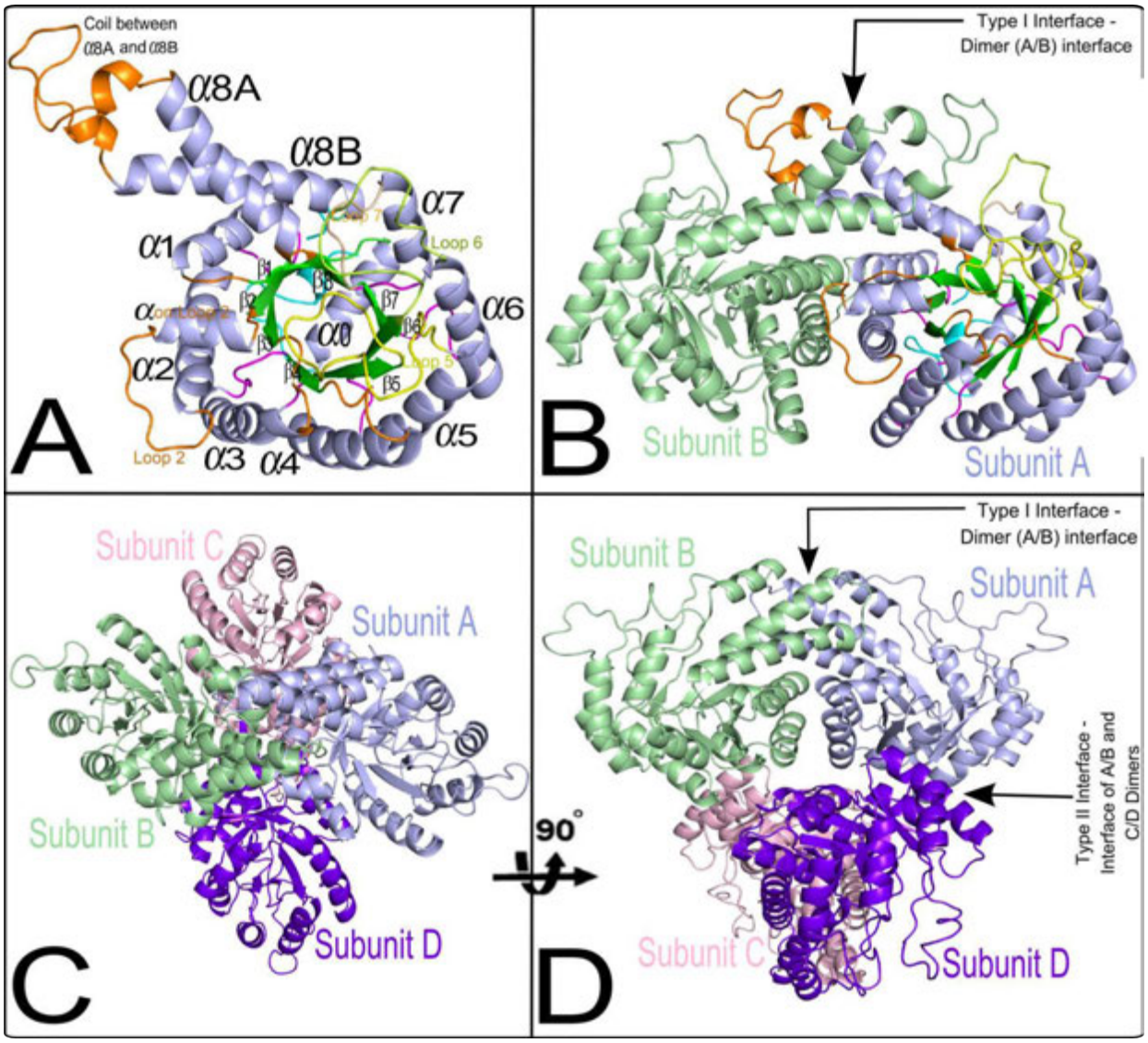


Figure 3. (A) Different structural components of a subunit of class II FBA (based on *E.coli* FBA with PDB Id 1B57). Eight strands β1 ~ β8 form the central barrel of the subunit and eight helices α1 ~ α8 surround the barrel. Helix α0 covers the bottom of the central barrel. Helix α8 has two segments – α8A and α8B that are connected by a coil between them. (B) A dimeric class II FBA (based on *E.coli* FBA with PDB Id 1B57). Two subunits (A and B) are joined by the interactions of a type 1 interface. The coil between the helices α8A and α8B of one subunit crosses over to come to the proximity of a front loop of the partner subunit. (C) A tetrameric class II FBA (based on *T.aquaticus* FBA with PDB Id 1RVG). Two type 1 dimers form the tetrameric structure through the interactions of two type 2 interfaces. Pair A & B join together with pair C & D through the interactions of two type 2 interfaces. (D) The tetrameric structure in panel C with a 90º bottom-up rotation.

Type I Interface Formation (residue indexing is based on *E.coli* FBA)

Figure 4 marks the components from each subunit that form the type 1 interface region. The structural components of a subunit of the dimer that are within 5Å distance are helices α1, α2, parts of helices α8A and α8B and the coil connecting them, front loop 2, and the tip of loop 6. These components from one subunit form complementary contact with the counterpart components from the partner subunit. The C-terminus of front loop 1 (residues 35:39) and the N-terminus of helix α1 (residues 40:52) of one subunit are buried by helix α1 and the helical structure within front loop 2 of the partner subunit upon interface formation. Also, helix α2 (residues 80:99 - GAAILGAISGAHHVHQMAEH) of one subunit runs anti-parallel to helix α2 of the other subunit as shown in Fig. 5A. These two helices are structurally complementary to one another as shown in Fig. 5B.

The front loop 2 of each subunit consists of the residues 61:80 and the N-terminus section of this loop has a small helical growth which formed by the residues 62:68. This helical part plays an important role in interface formation. The region from one subunit runs towards the other subunit where the second subunit forms a groove between helix 1 (residues 39:52) and helix 2 (residues 80:99) as shown in Fig. 5C. Symmetrically, the front loop 2 of the other subunit makes similar contacts with helix 1 and helix 2 of the partner subunit. Helices α8A and α8B of one subunit pair together and bind to the similar pair of helices α8A and α8B of the partner subunit as shown in Fig. 4. The front loop 6 from one subunit makes contact with the coil between helices α8A and α8B of the partner subunit. The tip (residues 229:231) of the front loop 6 of one subunit comes in close contact with the tip (residues 312:325 – especially, residues 312 – G, 313 – Q, 318 – K; 321 – D, and 325 – K) of the coil (residues 306:330) from the partner subunit.

The C-terminus of helix 8A and the N-terminus of helix 8B come in close contact with the C-terminus of helix 8B and the N-terminus of helix 8A of the partner subunit, respectively.

Together these structural components form a very strong type I interface in the FBA structures.

Type II Interface Formation (residue indexing is based on *T.aquaticus* class II FBA)

The N-terminal region of helix 0; the C-terminal region of helix 3; the C-terminal region of helix 4, back loop 4 and the N-terminal region of strand 5 contact similar regions on one of the partner subunits. Panel D of Fig. 3 shows a type II interface on a tetrameric *T.aquatics* FBA structure.

2.1.4 Sequence Conservation in *E.coli* FBA, *E.coli* TBA, and *T.aquaticus* FBA

The three structures selected for modeling the dynamics of aldolases are *E.coli* FBA, *E.coli* TBA, and *T.aquaticus* FBA. The set class II FBAs used in Fig. 6 lacks *E.coli* TBA, which has high subunit structural similarity with *E.coli* FBA but low sequence identity. However, functional *E.coli* FBA is a dimer but functional *E.coli* TBA is a dimer. Figure 6 shows the multiple sequence alignment of three class II aldolases – *E.coli* FBA, *E.coli* TBA, and *T.aquaticus* FBA. Table II catalogues all the residues that are conserved in these three aldolases. The residues that were also conserved in the larger set of 9 FBA sequences from Fig. 1 are marked as red. The set of aldolases used in Fig. 1 does not contain *E.coli* TBA. The difference between two the datasets used in Figs. 1 and 6 is the sequence of *E.coli* TBA, whose sequence conservation with the larger dataset used in Fig. 6 is very poor. Figure 6 gives a better picture of conservation between the three aldolases used in this research.

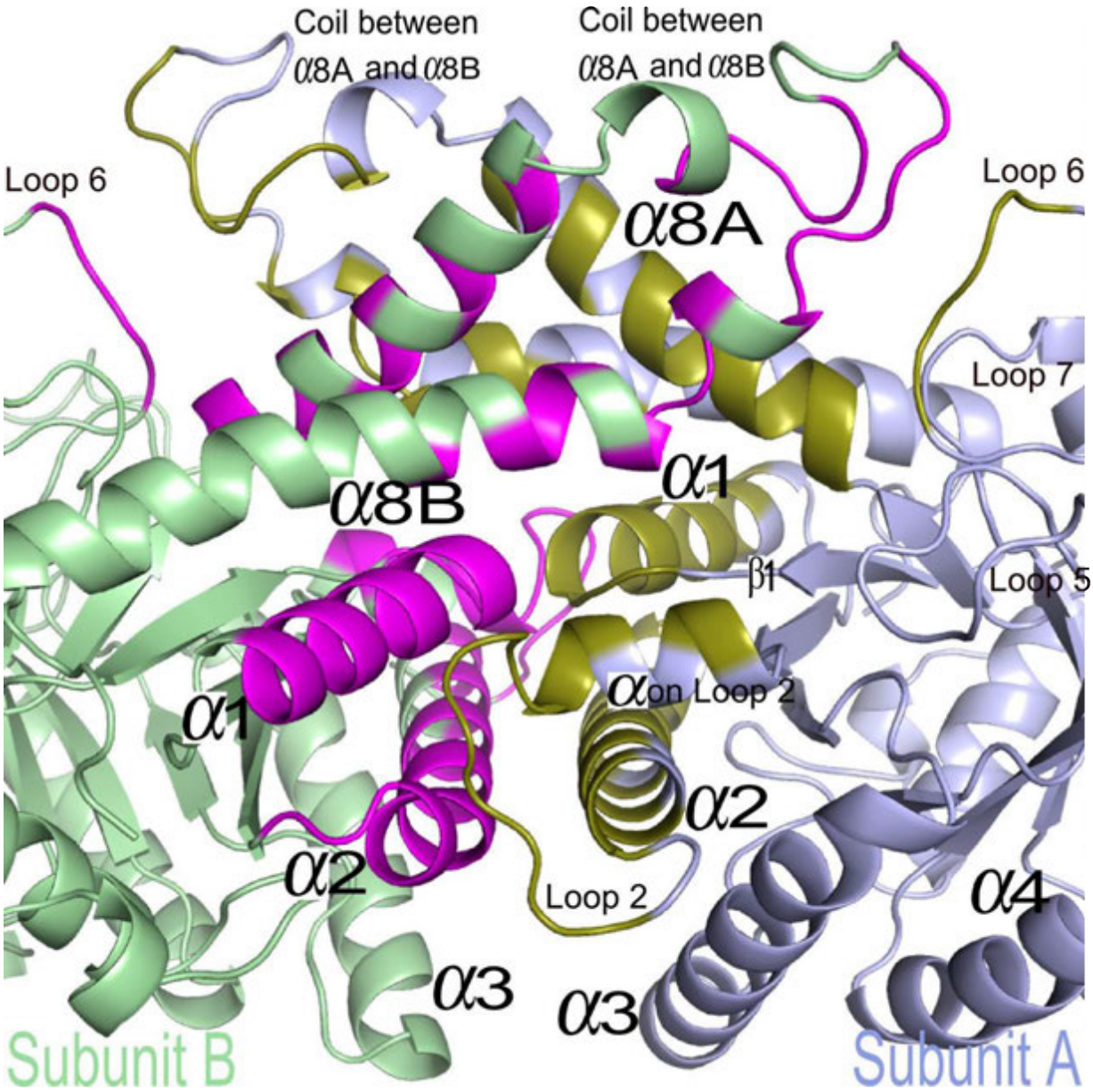


Figure 4. Components of type I interface of an *E.coil* FBA dimer. Green – interface forming residues in chain A that are within 5Å distance from chain B; Magenta – interface forming residues in chain B that are within 5Å distance from chain B;

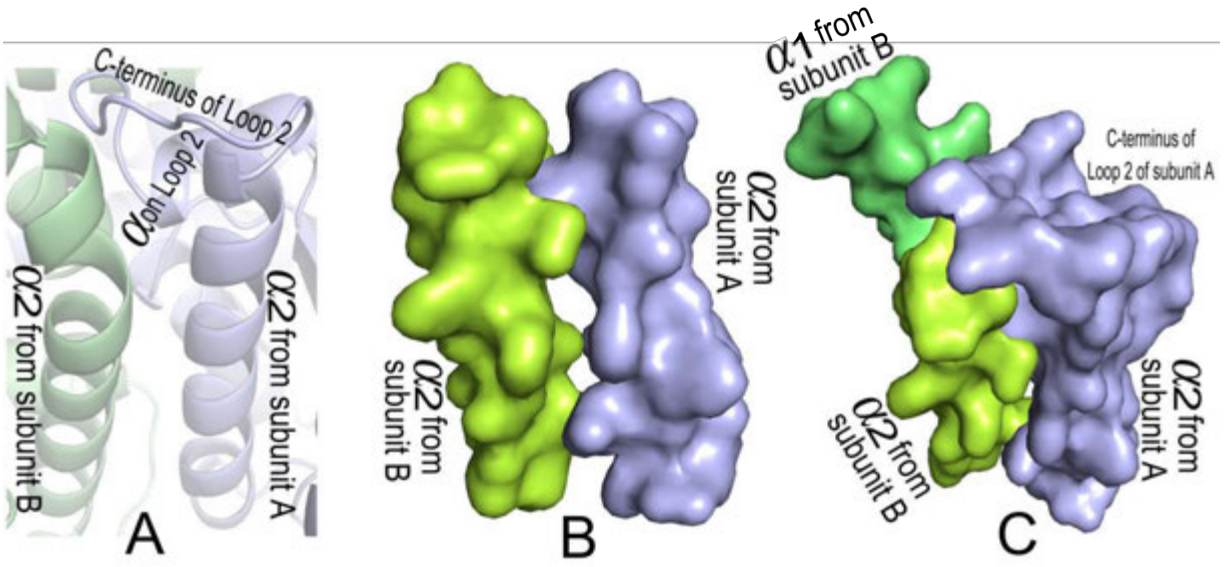


Figure 5. Type I interface complementarity (A) helix 2 from subunit A (light green) and helix 2 from subunit B (light blue) in anti parallel arrangement; (B) Helix 2A and helix 2B are complementary to each other; (C) The complementarity is further extended when the helical region of front loop 2 docks into the ridge between helix 1 and helix 2 of the partner subunit.

2.1.5 Formation of the Catalytic Site Microenvironment

Each subunit of an FBA structure has a catalytic site which is built from the following:

- The C-terminal ends of strand 7 and strand 8
- Front loop 6 and front loop 7
- C-terminus of helix 8A of the partner subunit
- Coil between helix 8A and helix 8B of the partner subunit

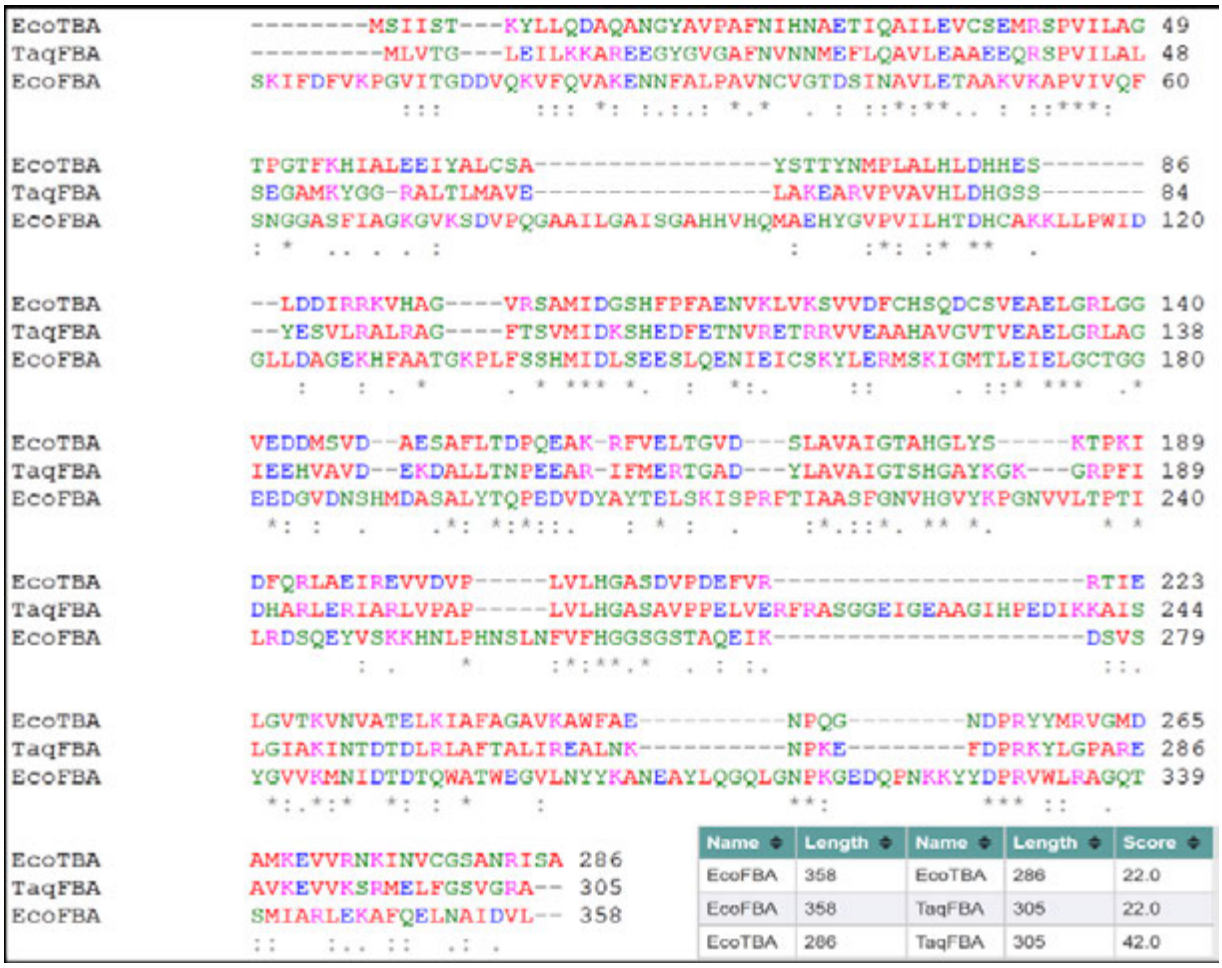


Figure 6. Multiple sequence alignment of three class II aldolase sequences – *E.coli* FBA, *E.coli* TBA, and *T.aquaticus* FBA. In the inset, the sequence alignment scores. The color code is the same as used in Fig. 1.

Catalytic environment in FBA structure consists of the following intersecting sets of residues:

- Set 1: The residues that hold the substrate and products in place as needed.
- Set 2: The residues to stabilize the Zn^{+2} ion in place as required in both the open and closed conformations of the enzyme.
- Set 3: The residues that do the catalysis.
- Set 4: The residues that are responsible to maintain the required motion of the catalytic loops.

Table III. Conserved Residues and their Locations in the Structure (*E.coli* FBA , *E.coli* TBA, *T.aquaticus* FBA) (Residues conserved across nine aldolases shown in Figs. 1 and 2 are in red)	
Residue (*E.coli* FBA, *E.coli* TBA, *T.aquaticus* FBA)	**Location on the Structure** (*E.coli* TBA, *T.aquaticus* FBA, *E.coli* FBA)
A (24, 13, 12)	helix 0
A (33, 22, 21); N (35, 24, 23)	strand 1
A (44, 33, 32); LE (46:47, 35:36, 34:35)	helix 1
PVI (55:57, 44:46, 43:45)	back loop 1 – strand 2
G (63, 52, 51)	front loop 2
P (103, 76, 74)	C-terminal end of back loop 2
H (107, 80, 78)	strand 3
DH (109:110, 82:83, 80:81)	(front loop 3, strand 3,strand 3)
A (132, 96, 94); S (140, 100, 98)	A-back loop 3, S-strand 4
MID (142:144, 102:104, 100:102)	strand 4/front loop 4
S (146, 106, 104)	front loop 4
N (153, 113, 111)	helix 4
E (172, 132, 130)	strand 5
ELG (174:176, 134:136, 132:134)	strand 5/front loop 5
G (180, 140, 138); E (182, 142, 140)	front loop 5
M/V/V (185, 145, 143)	front loop 5 in all aldolases
A (194, 152, 150)	front loop 5
T (197, 155, 153)	front loop 5
P (199, 157, 155); E (208, 165, 163); A (219, 173, 171)	P (helix 5, helix 5,front loop 5); E-helix 5, A-strand 6
G (223, 177, 175)	front loop 6
HG (226:227, 180:181, 178:179)	front loop 6
Y (229, 183, 181)	front loop 6
P (238, 187, 187); P (255, 204, 204); V (262, 206, 206)	P-front loop 6; P-back loop 6;V-strand 7
(HGAS,HGAS, HGGS – 264:267, 208:211, 208:211)	front loop 7
G (281, 225, 246)	back loop 7
K (284, 228, 249)	strand 8
N (286, 230, 251)	strand 8
T (289, 233, 254); A (294, 238, 259); NP (316:317, 250:251, 271:272); DPR (329:331, 255:257, 276:278)	T-front loop 8; A-helix 8A; NP and DPR – coil between helix 8A & helix 8B;

These four sets of residues in the three aldolases (*E.coli* FBA, *E.coli* TBA, and *T.aquaticus* FBA) and their locations on the structures are listed in Table IV and Table V. From these tables and Fig. 6, it is evident that most of the residues forming the catalytic microenvironment are well

conserved. Observing these residues in Fig. 1 and Table III will further reveal that these conserved residues are prevalent for a larger set of class II FBA proteins across the species.

Three main substrates/products for these enzymes are: FBP/TBP, DHAP, and GAP. Some of the residues that bind with these components are common to each structure. In *E.coli* FBA, Asn 35, Ser 61, and Arg 331 are involved in substrate binding together with other residues as described in Tables 4 and 5. Mutation of Asn 35 residue (N35A) reduces the enzyme activity to only 1.5% of the wild-type enzyme. Reactions also indicate that this mutation affects the binding of both triose substrates – DHAP and GAP [6]. It is located at the C-terminal end of strand 1. Mutation of Ser 61 residue to Ala increases the Km value for FBP by 16 fold which affects the enzyme's binding capability of GAP in the active site [6]. This is located at the C-terminal end of strand 2. Arg 331 is involved in binding of FBP in *E.coli* FBA; more precisely it interacts with the C-6 phosphate group of the substrate [7]. This is located on the C-terminal end of helix 8B of the partner subunit.

Three histidine residues and the substrate form the scaffold to hold the Zn^{+2} in place as shown in Fig. 7(B) [6]. In *E.coli* FBA, these three histidine residues are His 110, His 226, and His 264 located on front loops 3, 6, and 7, respectively. While the enzyme is in the open conformation, Glu 174, located on the C-terminal end of strand 5, is one of the ligands of the Zn^{+2} ion in its buried position [9]. Mutation of this residue (E174A) causes the enzyme's catalytic activity to be severely crippled, which implies that holding the Zn^{+2} ion in place in the catalytic environment in the open conformation is as important as it is in the closed conformation. Another residue Asp 144, located on the front loop 4, also is a ligand for Zn^{+2} ion [8] when the enzyme is in its open conformation. The Na^{+} binding site is about 5.6Å from the

Zn^{+2} binding site. This ion is sandwiched between front loops 6 and 7. The two cations help create the correct active-site alignment of the residues for the catalytic function [8].

Some of the residues are found to be of significant importance for catalysis. Mutation of any of Asp 109 and Asn 286 residues causes 300-fold and 800 fold decreases in the k_{cat} of the reaction[10] . Asp 109, which is located on front loop 3, aids binding of DHAP in the catalytic pocket. Because of the presence of a Zn^{+2} ion, some polarization of the ketone carbonyl group may facilitate the abstraction of proton from DHAP to generate the intermediate carbanion. Asn 286 located on the C-terminal end of strand 8 stabilizes the newly formed carbanion which then attacks the carbonyl of the incoming GAP to generate the new carbon-carbon bond for the condensation. Asp 109 is responsible for the polarization of the carbonyl group of GAP and can also donate a proton to stabilize the developing charge [10]. Mutation of Lys 325, which is located on the coil between helix 8A and helix 8B of the partner subunit, shows that it is more involved in catalysis than in binding. It may also play an indirect role to support the other important residues to form the catalytic micro environment [6]. Glu 181 and Glu 182, that lie on front loop 5, go through a large conformational change upon substrate binding and are placed in close proximity to the active site. Glu 182 functions in both directions in the FBA enzymatic activity – as a proton donor for aldolase cleavage and as a proton abstractor in the opposite condensation direction. A quadruple mutation of G176A, G179A, G180A, and G184A located on the same loop reduces the enzymatic activity significantly. This indicates that flexibility of the front loop 5 is important for proper functioning of the enzyme. Conservation of Gly 176, Gly 179, Gly 180, and Gly 184 across the species as shown in the MSA in Fig. 1 preserves the flexibility of this loop so that it can go through the required conformational switching between open and closed forms [9].

Table IV. Conserved Residues in three Class II Aldolases			
Residues that bind with the Substrate and Product	***E.coli* FBA**	***E.coli* TBA**	***T.aquaticus* FBA**
	DHAP	**DHAP**	**DHAP**
	Asp 109 – front loop 3	Asp 82 – strand 3;	Asp 80 – strand 3
	Glu 182 – front loop 5	Glu 142 – front loop 5	Glu 140 – front loop 5
	Asp 288 – front loop 8	Ala 232 – front loop 8 (Ala is smaller than Asp and provides extra space in the catalytic pocket)	Asp 253 – front loop 8
	Asp 329 – coil between helices 8A and 8B (from the partner subunit)	Asp 255 – coil between helices 8A and 8B (from the partner subunit)	Asp 276 – coil between helices 8A and 8B (from the partner subunit)
	GAP	**GAP**	**GAP**
	Asn 35 – strand 1	Asn 24 – strand 1	Asn 23 – strand 1
	Ser 61 – front loop 2	Thr 50 – front loop 2	Ser 49 – front loop 2
	Asp 109 – front loop 3	Asp 82 – strand 3	Asp 80 – strand 3
	Asp 288 – front loop 8	Ala 232 – front loop 8	Asp 253 – front loop 8
	FBP	**TBP**	**FBP**
	Asn 35 – strand 1	Asn 24 – strand 1	Asn 23 – strand 1
	Ser 61 – front loop 2	Thr 50 – front loop 2	Ser 49 – front loop 2
	Asp 109 – front loop 3	Asp 82 – strand 3	Asp 80 –strand 3
	Asp 288 – front loop 8	Ala 232 – front loop 8	Asp 253 – front loop 8
	Arg 331 – helix 8B (from the partner subunit)	Arg 257 – helix 8B from the partner subunit (from the partner subunit)	Arg 278 – helix 8B (from the partner subunit)

Table V. Conserved Residues in three Class II Aldolases				
	***E.coli* FBA**	***E.coli* TBA**	***T.aquaticus* FBA**	
Cation Binding Residues	**Zn^{+2}**	**Zn^{+2}**	**Co^{+2}**	
	His 110 – front loop 3	His 83 – strand 3	His 81 – strand 3	
	His 226 – front loop 6	His 180 – front loop 6	His 178 – front loop 6	
	His 264 – front loop 7	His 208 – front loop 7	His 208 – front loop 7	
	Asp 144 – front loop 4 (open conformation)	Asp 104 – front loop 4 (open conformation)	Asp 102 – front loop 4 (open conformation)	
	Glu 174 – strand 5 (open conformation)	Glu 134 – Strand 5 (open conformation)	Glu 132 – strand 5 (open conformation)	
	Na^{+}	**Na^{+}**	**Na^{+}**	**NH_4^{+}**
	Val 225 – front loop 6	Ala 179 – front loop 6	Ser 175 – front loop 6	His 78, Asp 80 – Strand 3;
	Gly 227 – front loop 6	Gly 181 – front loop 6	Gly 177 – front loop 6	Glu 130 – Strand 5;
	Tyr 229 – front loop 6	Tyr 183 – front loop 6	Tyr 179 – front loop 6	Asn 251 – Strand 8
	Gly 265 – front loop 7	Gly 209 – front loop 7	Gly 209 – front loop 7	**Y^{+}**
	Ser 267 – front loop 7	Ser 211 – front loop 7	Ser 211 – front loop 7	Asp 102, Ser 104 – front loop 4 Glu 132 – strand 5
Catalytic Residues	Asp 109 – front loop 3	Asp 82 – strand 3	Asp 80 –strand 3	
	Glu 174 – strand 5	Glu 134 – strand 5	Glu 132 – strand 5	
	Glu 181 – front loop 5	Val 141– front loop 5	Ile 139 – front loop 5	
	Glu 182 – front loop 5	Glu 142 – front loop 5	Glu 140 – front loop 5	
	Gly 265 – front loop 7	Gly 209 – front loop 7	Gly 209 – front loop 7	
	Asn 286 – strand 8	Asn 230 – strand 8	Asn 251 – strand 8	
	Asp 288 – front loop 8	Ala 232 – front loop 8	Asp 253 – front loop 8	
	Lys 325 – coil between helices 8A and 8B (from partner subunit)	No equivalent of *E.coli* FBA Lys 325	No equivalent of *E.coli* FBA Lys 325	
	Asp 329 – coil between helices 8A and 8B (from the partner subunit)	Asp 255 – coil between helices 8A and 8B (from the partner subunit)	Asp 276 – coil between helices 8A and 8B (from the partner subunit)	

It is evident from Tables 4 and 5 that these catalytically important residues are mostly conserved in *E.coli* TBA [11] and *T.aquaticus* FBA [12]. There are few significant differences in *E.coli* TBA and *T.aquaticus* FBA that set them apart from *E.coli* FBA. First, though *T.aquaticus* FBA has Asp 253, which is equivalent to *E.coli* Asp 288, *E.coli* TBA has Ala 232 in the same position. Absence of this Asp in *E.coli* TBA gives more room for the substrate to adopt multiple conformations thus making the enzyme suitable for catalysis of TBP. Second, neither *E.coli* TBA nor *T.aquaticus* FBA has a residue equivalent to Lys 325. The role of Lys 325 in *E.coli* FBA is thought to be more supporting of other catalytically important residues to form the catalytic environment rather than playing a more direct role in catalysis, probably the need for such a residue is eliminated because of the shortening of the coil region between helix 8A and helix 8B in *E.coli* TBA and *T.aquaticus* FBA. Another difference in hyperthermophilic *T.aquaticus* aldolase is that it is a Co^{+2} based enzyme and the monovalent cation is either NH_4^+ or Y^+(Yttrium) [11;12] .

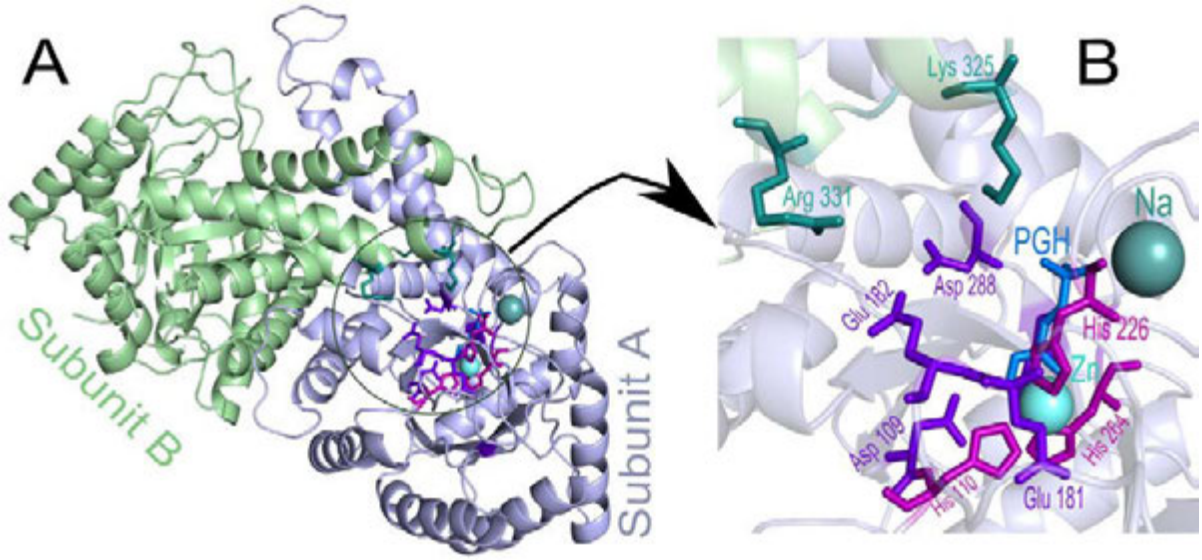


Figure 7. (A) Location of catalytic environment in subunit A of the class II FBA of *E.coli* is circled. (B) Some of the key residues, substrate, Zn^{+2} and Na^+ ions involved in the catalysis are labeled in the zoomed figure. The carbon atoms of Arg 331 and Lys 325 from the partner subunit are in cyan [6].

2.2 Results

2.2.1 Oligomerization and Stability at the Interface Regions

Front Loop 2

The length of this loop is different in the three structures. It is 19 residues long in *E.coli* FBA, and in *E.coli* TBA and *T.aquaticus* FBA, it is 9 and 10 respectively, as shown in Fig. 8. This loop is important in forming the type I interface in each of the structures and establishing the functional dynamics of the oligomeric aldolase structures. A helical region on this loop forms a beam like component that docks onto the ridge between helix 1 and helix 2 of the partner subunit of type I dimer. The length of this helical region on the loop can vary across these three structures: *E.coli* FBA 8 (62:69), *E.coli* TBA 4 (53:56), *T.aquaticus* FBA 7 (50:56). The helical part and the C-terminal region of front loop 2 are stabilized upon dimerization in each case. They are stabilized even further upon tetramerization in *E.coli* TBA and *T.aquaticus* FBA as shown in panels C and D of Fig. 9.

Helix 2

It is an important component of the structure. It plays a role in forming the type I interface during dimerization. Helix 2 of one subunit comes in contact with helix 2 from the partner subunit in anti-parallel complementary alignment as shown in Fig. 4 and Panel A of Fig. 9. Dimerization stabilizes the motion of this region by reducing its fluctuation in each of the structures as shown in panels B, C, and D of Fig. 9.Tetramerization in *E.coli* TBA and *T.aquaticus* FBA further stabilizes its fluctuation as shown in panels C and D. This indicates that formation of a tetramer through the interactions along the type 2 interface, allosterically lowers

the fluctuation along this region. Because of this lowered fluctuation helix 2 region, tetramerization gives brings more stability to the structure.

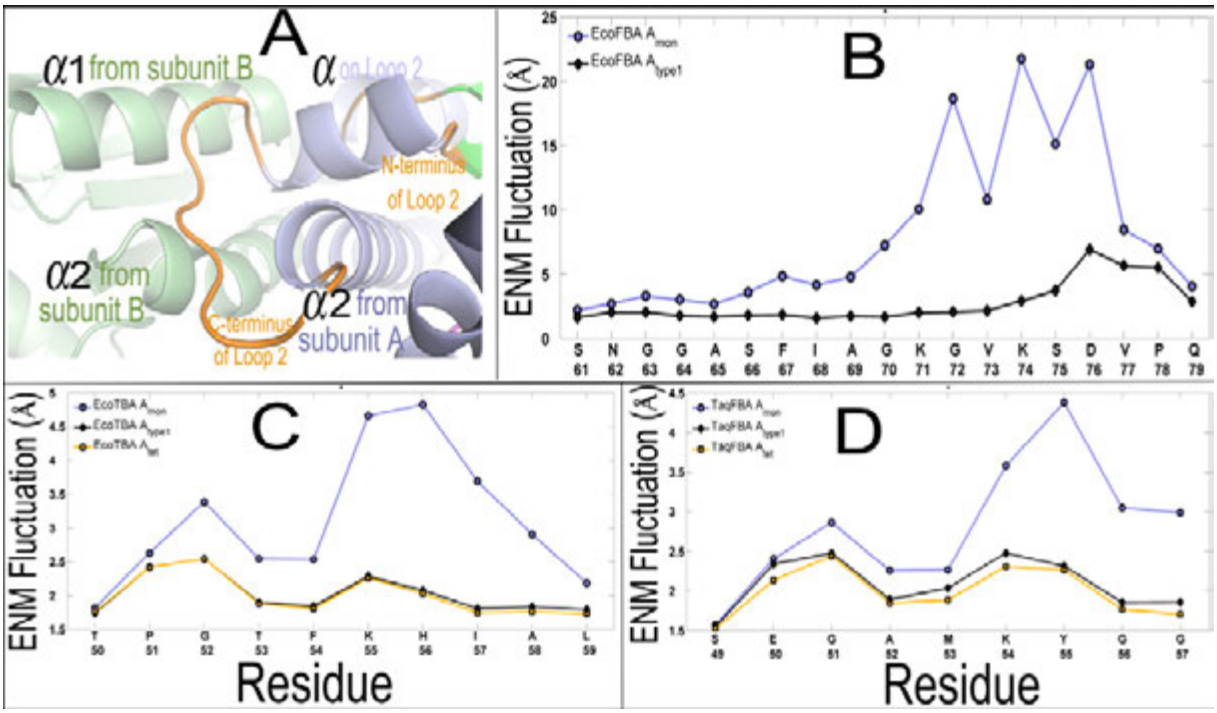


Figure 8. Fluctuations of front loop 2. The residue indices for this loop region are: *E.coli* FBA – 61:79; *E.coli* TBA – 50:59; and *T.aquaticus* FBA – 49:57. The helical segment on the N-terminus of the loop consists of the residues – *E.coli* FBA – 62:69; *E.coli* TBA – 53:56; and *T.aquaticus* FBA – 50:56. (A) The beam like helical region on front loop 2 of one subunit docks between helices α1 and α2 of the partner subunit. (B) The fluctuations of the loop front loop 2 in monomer and dimer of *E.coli* FBA. (C) The fluctuations of the loop front loop 2 in monomer, dimer, and tetramer of *E.coli* TBA. (D) The fluctuations of the loop front loop 2 in monomer, dimer, and tetramer of *T.aquaticus* TBA.

Helix 8A, helix 8B, and the coil that connects them

Helix 8A and helix 8B and the coil connecting them are important, structurally and functionally. The length of the coil gets is much shorter in the tetrameric aldolase structures (*E.coli* TBA and *T.aquaticus* FBA) compared to the dimeric aldolase structure (*E.coli* FBA): *E.coli* FBA 25 (306:330), *E.coli* TBA 7 (250:256), *T.aquaticus* FBA 6 (271:276). The helices take part in the formation of type I interface as shown in panels A and B of Fig. 5. This region

contains the catalytically important residues – Asp 329 and Arg 331 in *E.coli* FBA; Asp 255 and Arg 257 in *E.coli* TBA; and Asp 276 and Arg 278 in *T.aquaticus* FBA. The tip of the coil from one subunit comes in close proximity to one of the catalytic loops of the partner subunit in *E.coli* FBA (front loop 6) and *T.aquaticus* FBA (front loop 7). This stabilizes the fluctuation of helix 8A, helix 8B and the coil connecting them. Though in *E.coli* TBA, the coil does not come as near to any of the functional loops, still these regions are stabilized as shown in Fig. 10B.

2.2.2 Oligomerization and Functional Loop Motions around the Catalytic Microenvironment

In both *E.coli* FBA and *T.aquaticus* FBA dimerization, because of the proximity of the tip of the coil between helix 8A and helix 8B to one of the functional loops – front loop 6 in *E.coli* and front loop 7 in *T.aquaticus*, the fluctuations of the contacting loop decrease as evident from the ENM fluctuations as shown in panels A and C of Fig. 10. However, in case of *E.coli* TBA, the coil does not come close to any of the functional loops; thus, oligomerization in this enzyme allosterically increases the motions of all of the functional loops.

Figure 11A shows that, in *E.coli* FBA, the flexibility of front loop 7 is constrained because of the spatial proximity of this loop and the coil between helix 8A and helix 8B of the partner subunit as depicted in panel B of the figure. Figure 11C marks the residues of the microenvironment on different functional loop regions. Figure 12 shows an enlarged view of this site – panel A showing the substrate/product sub-region and panel B showing the metal binding site.

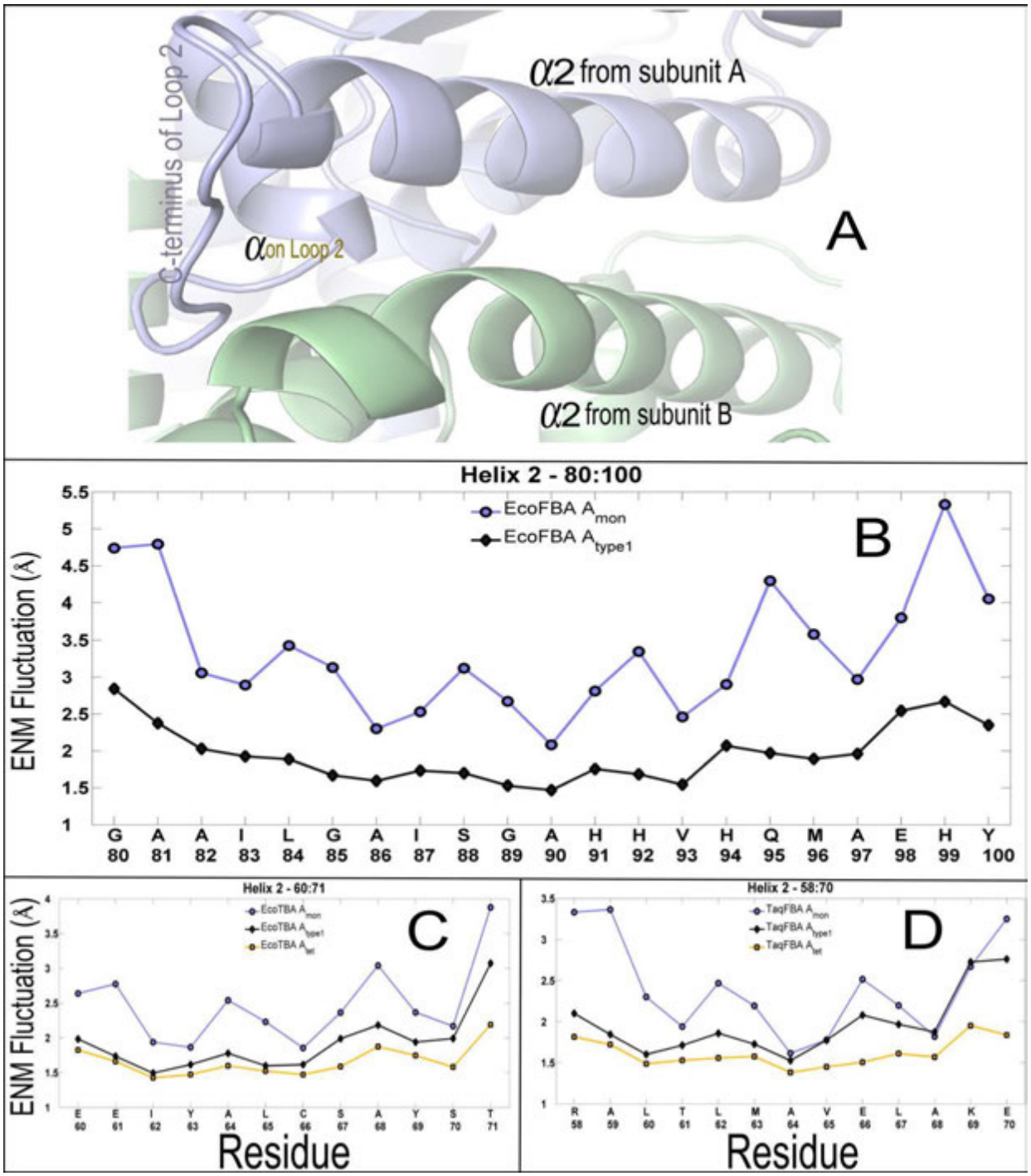


Figure 9. Fluctuations of helix 2. The residue indices for this segment in the three aldolases are: *E.coli* FBA - 80:100; *E.coli* TBA – 60:71; *T.aquaticus* FBA – 58:70. (A) Helix α2 from subunit A pairs with helix α2 from the partner subunit B. (B) Fluctuations of helix α2 from *E.coli* FBA; (C) Fluctuations of helix α2 from *E.coli* TBA; (D) Fluctuations of helix α2 from *T.aquaticus* FBA.

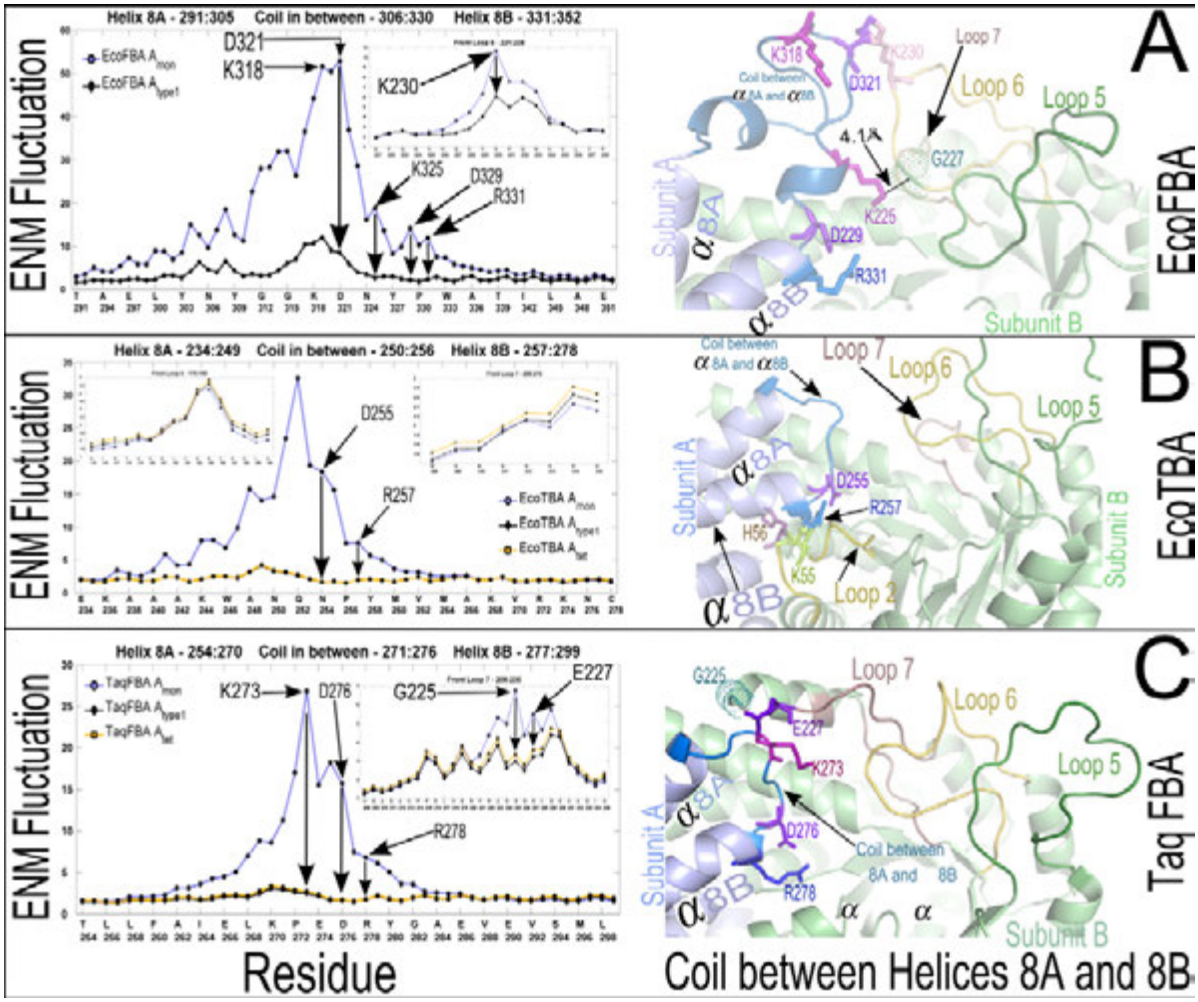


Figure 10. Changes in the fluctuations of helix 8A, helix 8B, and the coil between them. (A) (right) contact between front loop 6 and the coil connecting helix 8A and helix 8B of the partner subunit in *E.coli* FBA; (left) fluctuations of the helix 8A, helix 8B, and the coil; fluctuation of front loop 6 (in set); (B) (right) (right) No contact between front loop 6 and the coil connecting helix 8A and helix 8B of the partner subunit in *E.coli* TBA; (left) fluctuations of helix 8A, helix 8B, and the coil; fluctuation of front loops 6 and 7 (in set); (C) (right) contact between front loop 7 and the coil connecting helix 8A and helix 8B of the partner subunit in *T.aquaticus* FBA; (left) fluctuations of the helix 8A, helix 8B, and the coil; fluctuation of front loop 7 (in set);

The residues Lys 325, Asp 329, and Arg 331 located on helix 8B of the partner subunit take part in forming the catalytic environment. Figure 10A shows that these residues are stabilized upon dimerization –the relative magnitude of the Lys 325 attenuation being the largest. Lys 325 also makes a large swing from the enzyme's open to closed conformation. Panel C and panel D of Fig. 12 show the spatial rearrangement of Lys 325 residue in the open and closed

conformations. The distance between the substrate (DHAP) and Lys 325 is 17.9Å and 5.9Å between the open and closed conformations, respectively. The tip of this residue makes an excursion of 13.9Å whereas the base makes a 4.6Å shift, both between the open and closed conformations. Combined with the dimerization attenuation and conformational switch from open to closed, this residue is identified as an important part of the catalytic microenvironment in a dimeric *E.coli* FBA structure.

The presence of the two positive ions, divalent Zn^{+2} and monovalent Na^{+}, is important for catalysis. The Zn^{+2} ion binding residues His 110, His 226, His 264, Asp 144, and Glu 174 are stabilized upon dimerization as shown in panels A, E, G, and H, respectively of Fig. 11. The Na^{+} binding residues Val 225, Gly 227, Tyr 229, Gly 265, and Ser 267 are stabilized upon dimerization as well. The fluctuations of these residues are marked in panels A, E, and G.

The gray shaded oval shapes in panel E, F and G, of Fig. 11 mark the stabilization of the catalytic residues Asp 109, Glu 174, Glu 181, Glu 182, Gly 265, and Asp 288. Asp 288 is important for both catalysis and substrate binding. Not only is it involved in catalysis, but it also maintains the volume that needed to allow the substrate to sample its conformational space. Two other catalytic residues Lys 325 and Asp 329 from helix 8B of the partner subunit are also stabilized upon dimerization as shown in Fig. 10A.

Figure 13C shows the construction of the catalytic cavity of tetrameric *T.aquaticus* FBA. This consists of components from both subunits joined by a type I interface. It is important to note that front loop 7 comes in close proximity with the coil between helix 8A and helix 8B of the partner subunit – the residues making the contacts seen in this figure: Glu 227 from the loop region and Lys 273 from the coil. This suppresses the motions of the front loop 7 as shown in Fig. 13A. This also attenuates the mobility of Asp 276 and Arg 278, two substrate binding

residues contributing from the partner subunit. Fluctuation along other functional loop regions increases with dimerization. Tetramerization does not increase this as much anymore compared to dimerization.

The structure of the catalytic pocket in tetrameric *E.coli* TBA is different from that of *E.coli* FBA or *T.aquaticus* FBA is organized in such a way that none of the functional loops from one subunit come in close contact with any part of the partner subunit. Figure 14C shows such a microenvironment of the catalytic pocket of *E.coli* TBA. It is evident from this figure and Table II that the shortened length of coil between helix 8A and helix 8B of the partner subunit is unable to make such a contact with either functional loop 6 or 7. The consequence of this architectural difference is that oligomerization increases the fluctuations of all functional loops in this case. However, the catalytically conserved residues maintain a similar spatial arrangement. The distance between Asp 255 and GAP is 2.4 Å and the distance between Arg 257 and GAP is 7.4 Å. Catalytically important residues Val 141 and Glu 142 are on front loop 5. Divalent Zn^{+2} ion is held by His 83, His 180, His 208 and the substrate. Asp 104 (on front loop 4) and Glu 134 (on front loop 5) function as ligands for the Zn^{+2} ion in the open conformation. Panels (A), (B), and (D) of Fig. 14 show how ENM captures this enhanced dynamics of the functional loops upon oligomerization.

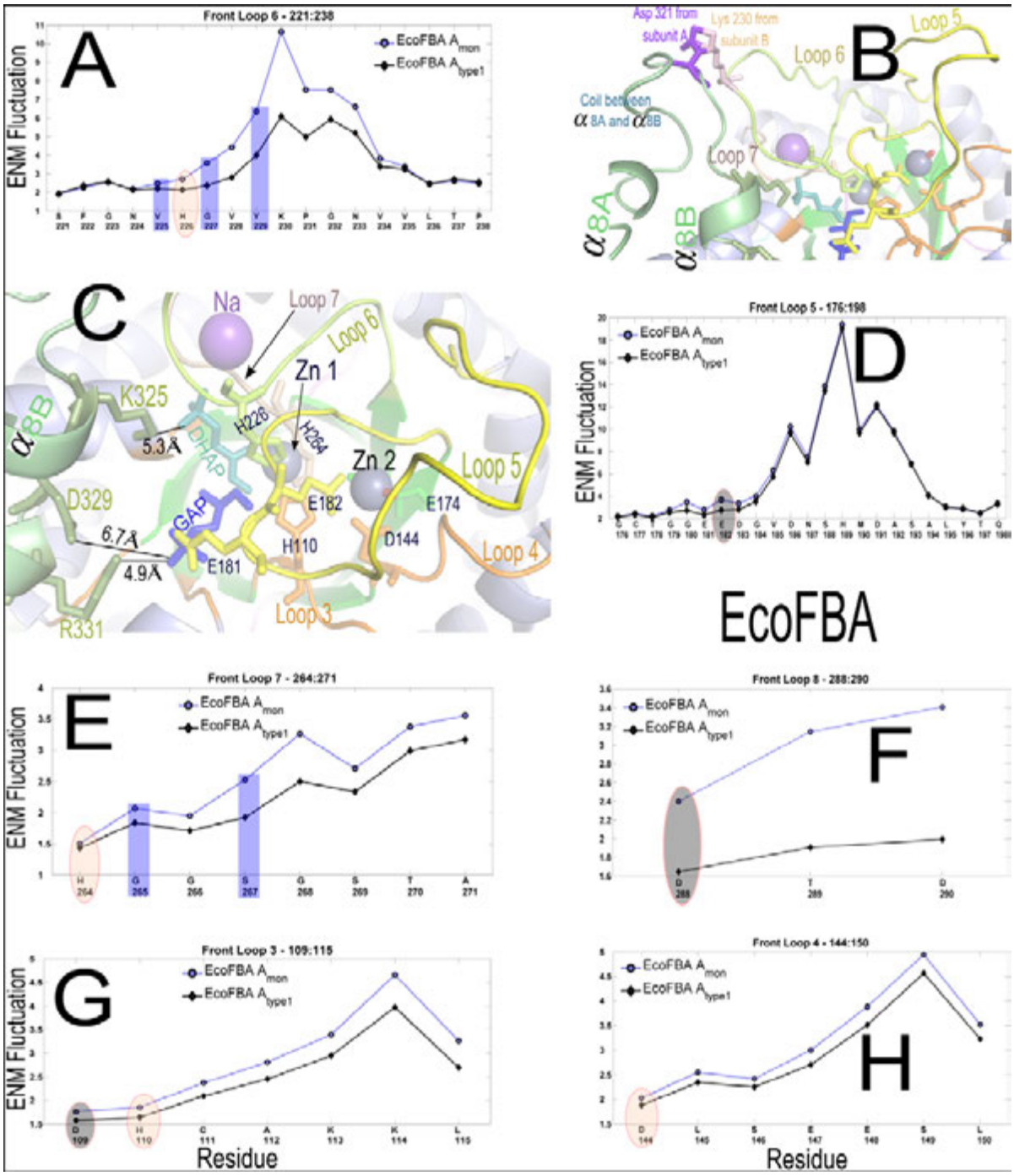


Figure 11. Fluctuations of functional loops in three aldolases. (A) Fluctuations of front loop 6; (B) Contact between coil and front loop 6; (C) The catalytic environment; (D) Fluctuations of front loop 5; (E) Fluctuations of front loop 7; (F) Fluctuations of front loop 8; (G) Fluctuations of front loop 3; (H) Fluctuations of front loop 4. Substrate binding residues are marked with gray ovals. Zn^{+2} ion binding residues are marked with pink ovals. Na^{+} binding residues are marked with blue rectangles.

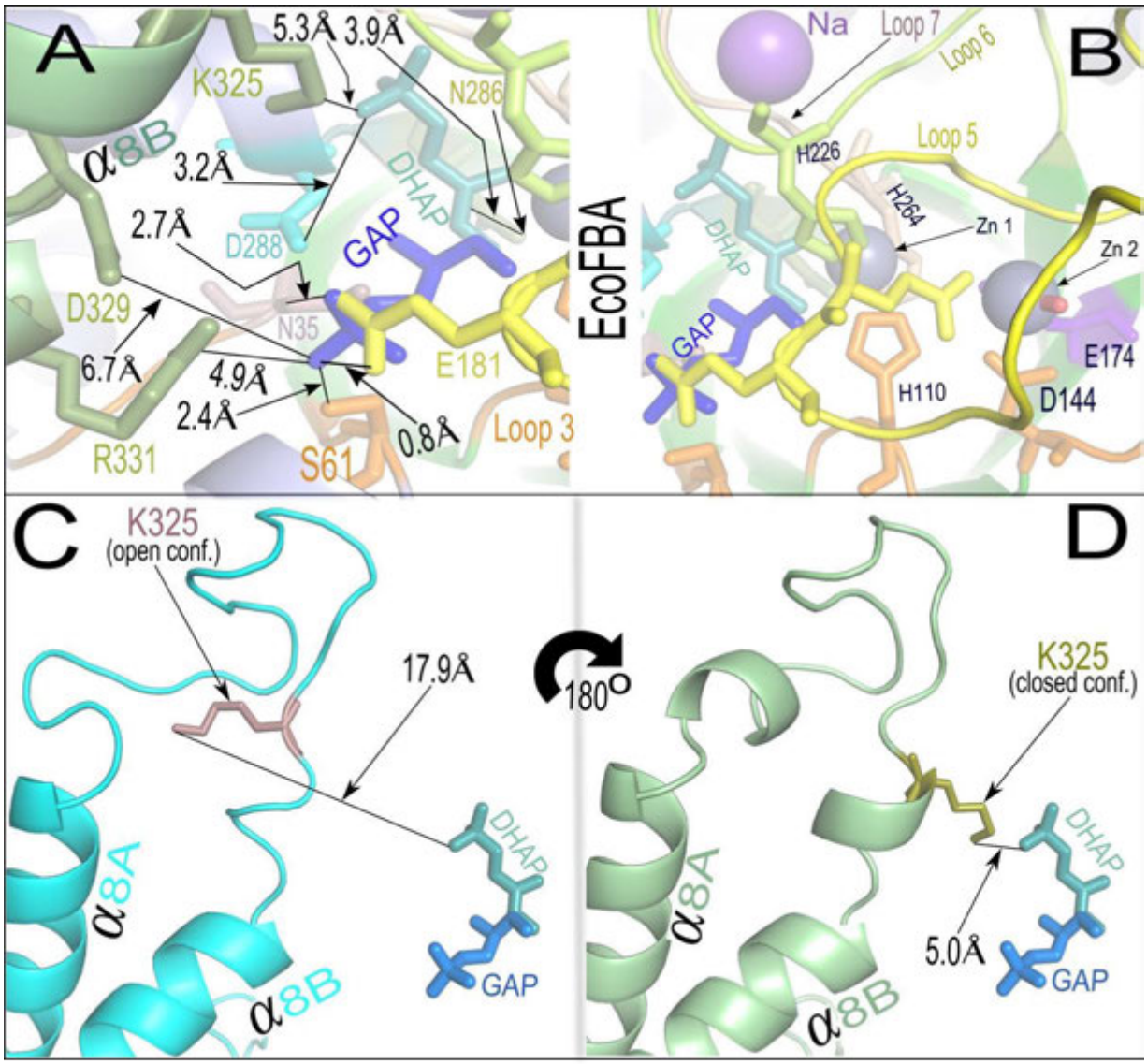


Figure 12. Microenvironment of aldolase catalytic pocket. (A) Substrate binding site of the catalytic microenvironment; (B) metal binding site if the catalytic microenvironment – ; (C) helix 8A, helix 8B, and coil region between them in open conformation – PDB Id 1ZEN; (D) helix 8A, helix 8B, and coil region between them in closed conformation – PDB Id 1B57; Large swing of Lys 325 from open to closed conformation brings it to the catalytically favorable distance from the substrate.

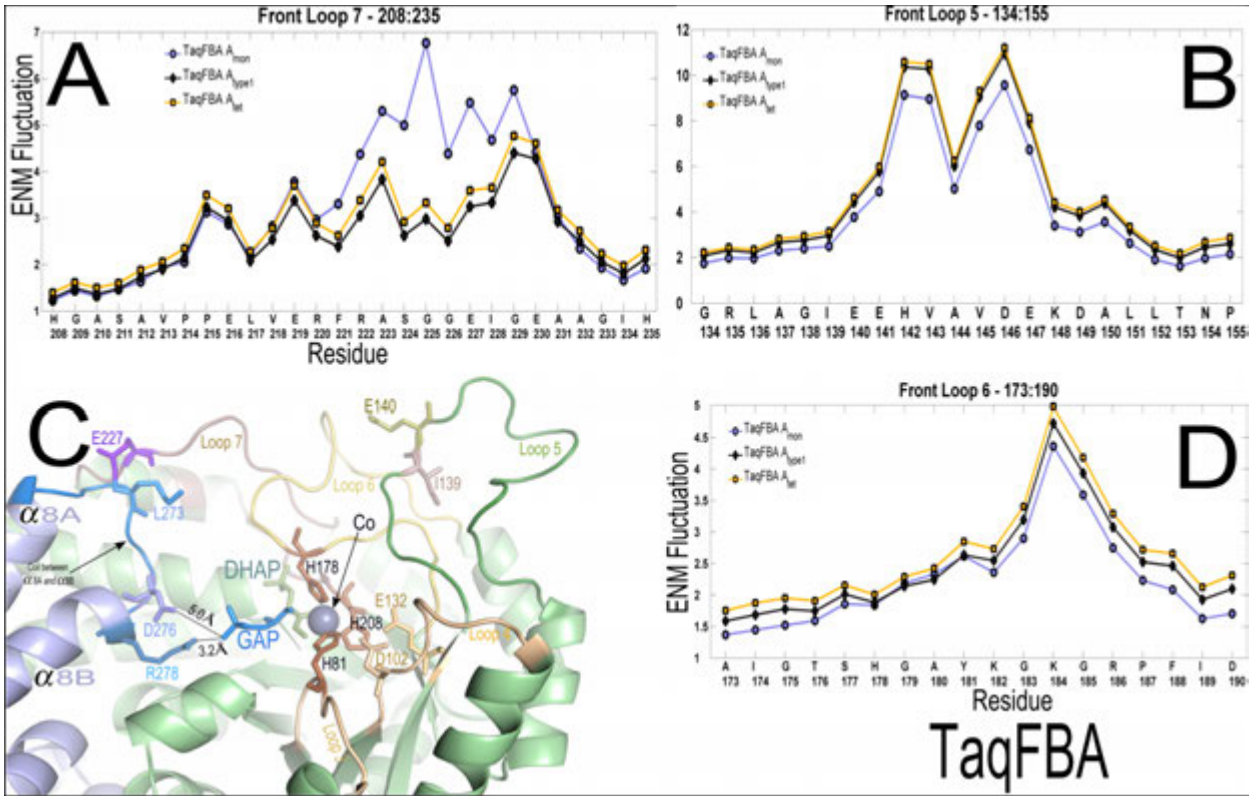


Figure 13. Change of fluctuations of functional loops 5, 6, and 7, with oligomerization in *T.aquaticus* FBA. (A) Change of fluctuations of front loop 7; (B) Change of fluctuations of front loop 5; (C) The catalytic microenvironment of *T.aquatiucs* FBA; (D) Change of fluctuations of front loop 6.

2.3 Discussion and Conclusion

This research has attempted to answer two questions regarding class II aldolase oligomerization:

- How does the oligomerization impact the stability of the structure?
- Does this also affect the functionality of this protein? If so, how?

Dimerization through a type I interface affects the global and local motions of the protein. The interface has two primary components – (1) helix 8A and helix 8B and the coil between them. The coil can form a contact with front loop 6 (in case of *E.coli* FBA) or front loop 7 (in case of *T.aquaticus* FBA) or it may not form any contact with any of the functional loops (in case of *E.coli* TBA) of the partner subunit. Regarless of whether this contact is present or not,

this interface stabilizes helices 8A and 8B and the coil between them. When the contacts occur, this also reduces the fluctuations of the functional loop (front loop 6 in *E.coli* FBA and front loop 7 in *T.aquaticus* FBA) that come into proximity with the coil from the partner subunit and this attenuation propagates to the other functional loop regions in *E.coli* FBA structure. In two other cases, dimerization increases the fluctuations of the functional loops 5 and 6. (2) (a) Helix 2 forms an anti-parallel complementary contact with helix 2 of the partner subunit and (b) front loop 2 is placed against the ridge formed by helix 1 and helix 2 of the partner subunit – this arrangement is shown in Fig. 5. This interdigitation of front loop 2 with helix 1 and helix 2 of the partner subunit stabilizes the interface components: front loop 2 and helix 2, and also helix 1 and helix 2 of the partner subunit. Stabilization of front loop 2 is functionally important.

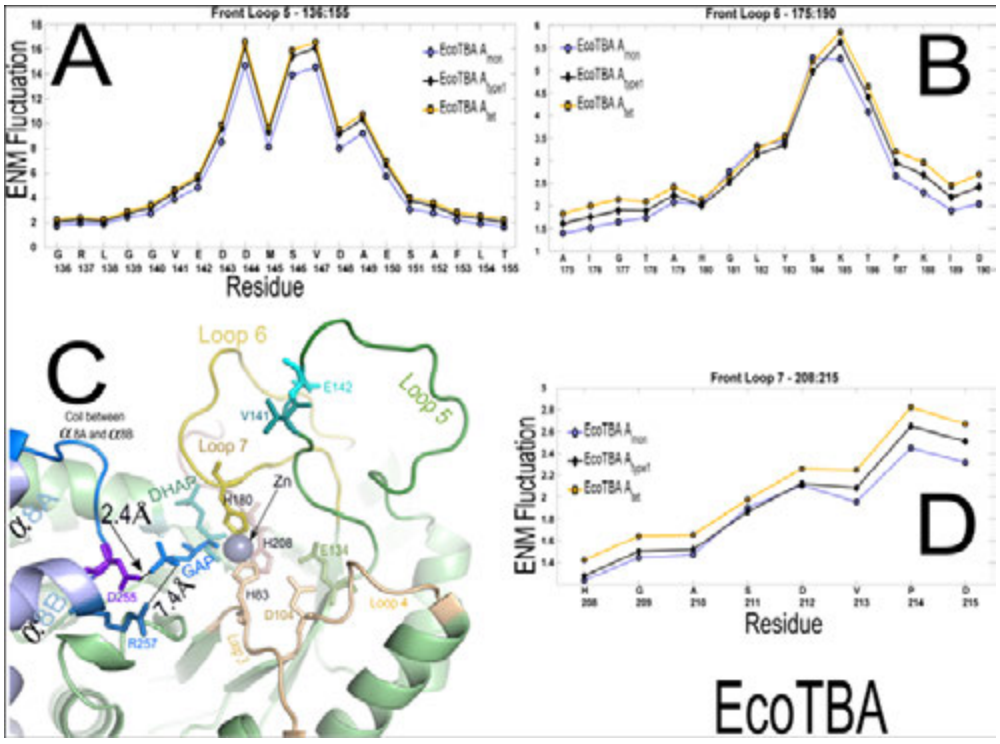


Figure 14. Change of fluctuations of functional loops in *E.coli* FBA with oligomerization. (A) Fluctuations of front loop 5 of monomer, dimer and tetramer; (B) Fluctuations of front loop 6 of monomer, dimer, and tetramer; (C) The catalytic microenvironment; (D) Fluctuations of front loop 7 of monomer, dimer, and tetramer.

2.4 Materials and Methods

2.4.1 Dataset Preparation

Class II FBA Structures for MSA and RMSD Calculation

For RMSD calculation of the FBA structures for the nine organisms that are used in the multiple sequence alignment, we use subunits from the following PDB Ids: *B.anthracis* – 3Q94; *C.immitis* – 3PM6; *C.jejune* – 3QM3; *E.coli* – 1B57; *G.lamblia* -2ISV; *M.tuberculosis* – 3EKL; *S.cerevisiae* – Model generated using I-TASSER using 1B57 as a template; *T.aquaticus* – 1RVG; *T.caldophilus* – 2FJK.

Aldolase Structures for Modeling Dynamics

We selected the following three PDB structures to investigate their dynamics – *E.coli* FBA (EcoFBA) – PDB Id:1B57; *E.coli* TBA (EcoTBA) – PDB Id:1GVF; *T.aquaticus* FBA (TaqFBA) – PDB Id:1RVG.

Obtaining Proper Oligomeric Forms

We use the symmetry information encoded in the PDB file to obtain the correct oligomeric forms for the structures selected for dynamics investigations. Also, highly mobile functionally essential loop regions (loops 6 and 7 in FBA) cannot be resolved by X-ray crystallography. We use the loop modeling program Modeler [13] of the Modeler software suite [13-15] to model the missing loop regions of the FBA structures retrieved from the Protein Data Bank [16].

E.coli **FBA: PDB Id – 1ZEN**. This PDB structure is also a dimer in its open conformation from *E.coli*, a mesophilic organism [17]. The structure has 17 missing residues (177 ~ 193) on loop 5, which is a highly extended loop of 25 residues. We have used loop modeling tools from

Modeler to model this loop. **PDB Id – 1B57**. This PDB structure is a dimer in its closed conformation from *E.coli*, a mesophilic organism [8]. It also has 12 missing residues (183 ~ 194) on loop 5 and in the same way as in the case of 1ZEN, we have modeled this missing loop segment.

E.coli **TBA: PDB Id – 1GVF**. This is a tetrameric TBA structure from *E.coli* which is a mesophilic organism [11]. It catalyzes the reversible cleavage of TBP into DHAP and GAP. The PDB submission is a closed dimeric structure (chain A and chain B) with symmetry information available to produce a tetramer. Chain A and chain B have two missing parts - 11 (140 ~ 150) and 8 (142 ~ 149) missing residues, respectively. These are functional loop 6 regions. We modeled the missing residues by using Modeler [13-15]. Then,by using the *sym exp sym* command of Pymol software [18], that utilizes the symmetry information, we produce a tetrameric structure for this structure.

T.aquaticus **FBA: PDB Id – 1RVG**. This is a tetrameric FBA structure from *T. aquaticus* that is an extreme thermophile [12]. In PDB database, 1RV8 has two biological units – bu1 and bu2. They have 8 missing residues in chain A (residues 140 ~ 147). This region happens to fall within one of the functional loops, which is front loop 6. Considering the importance of this loop, we modeled this missing segment by the loop modeling tool of the Modeler program [13]. 1RVG has two dimeric biological units – bu1 (chain A and chain B) and bu2 (chain C and chain D). Chain A is in the closed conformation and the others are in open conformation. Also chain A has six missing residues (142 ~ 147). We modeled these missing residues by using Modeler. The tetramer structure is retrieved by using the symmetry information encoded in the PDB file. Each chain of the tetramer thus retrieved is in the open conformation.

Comparing FBP and TBP

FBP and TBP have the same molecular formula, $C_6H_{14}O_{12}P_2$ with the same 2D representation shown in Fig. 15A, and the same molecular weight, 340.12, [19;20]. However, they have different 3D conformations as shown in panel B and panel C of Fig. 15. Also, each of them samples a different conformational space.

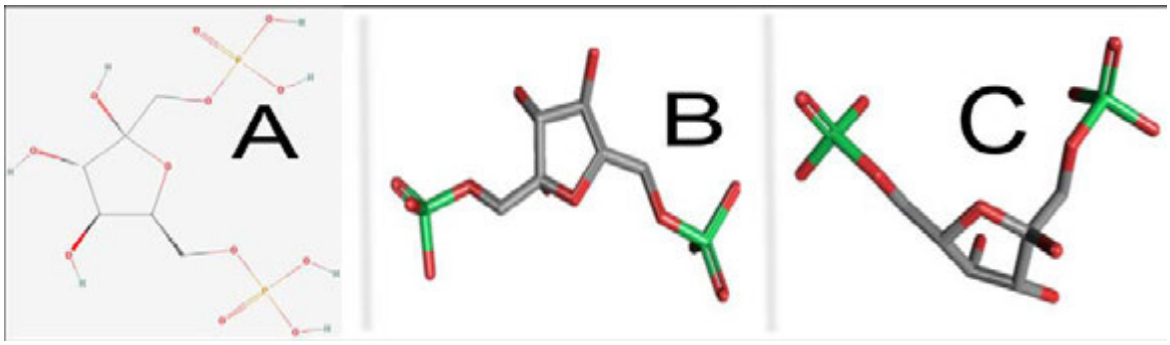


Figure 15. (A) The same 2D representation of FBP and TBP; (B) 3D representation of FBP (http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=10267); (C) 3D representation of TBP (http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?sid=6535&viewopt=PubChem) [19;20].

Depicting Catalytic Microenvironment

There is no aldolase structure in PDB with both of the components DHAP and GAP bound in the catalytic pocket for *E.coli* FBA, *E.coli* TBA, or *T.aquaticus* FBA. In each structure of *E.coli* FBA (PDB Id 1B57), *E.coli* TBA (PDB Id 1GVF), and *T.aquaticus* FBA (PDB Id 1RVG), we cut the substrates GAP and DHAP from 3EKZ (PDB structure for *M.tuberculosis* class II FBA) to place it into the respective catalytic pocket. The RMSD between one subunit of 3EKZ and one subunit of each of these structures are: *E.coli* FBA – 0.89Å, *E.coli* TBA – 1.46Å, and *T.aquaticus* 1.23Å. The small RMSD difference between a 3EKZ subunit and a subunit of each of the other structures gives us confidence that the placement of the substrates in the microenvironment as shown in panel C of Figs. 11, 13, and 14, is acceptably accurate for

depicting the distance relationships between the substrates and the functional residues in the microenvironment.

2.4.2 Modeling *S.cerevisiae* FBA Structures

There is no aldolase structure for the *S.cerevisiae* FBA in the PDB database. We use the homology based structure modeling component of the I-TASSER software suite [21;22] to build a structural model of yeast FBA. The homology based modeling of I-TASSER is a template guided modeling process. It has four stages. In stage 1, the program identifies a set of template structures from the PDB based on the query sequence. In stage 2, a set of continuous fragments are generated from the template structures. These are used to assemble a set of structural conformations of the sections of the query sequence that align well with the fragments. The sections that do not align well, usually loops/tails, arer then constructed by using *ab initio* modeling [23;24]. This set of conformations is clustered and then the cluster centroids are obtained by averaging the 3D coordinates of the structure in each cluster. In stage 3, another round of fragment assembly simulation is performed using the selected cluster centroids found in the previous stage. In the final stage, the accuracy of the predicted model is calculated. The c-score for accuracy is calculated based on the quality of the threading and convergence of structural assembly refinement in stages 1 and 2, respectively.

We obtain the sequence of *S.cerevisiae* from yeastgenome.org. We use the *E.coli* FBA structure as a template to guide the modeling process in I-TASSER. The sequence identity between *S.cerevisiae* and *E.coli* FBA is 48%. Hence, *E.coli* FBA structure can be considered to be a very good template – any template above 30% identity is usually a good template. I-TASSER returns a model with a high c-score, 1.03. C-score is in the range of [-5,2], with a more

positive value indicating a better structure. We accept our model with c-score 1.03 as a very good one.

2.4.3 Modeling Dynamics

To model the dynamics of the protein structure, we use the Anisotropic Network Model (ANM) [25]. To apply the ANM method to model the dynamics of a protein structure, we coarse-grain the structure where each residue is represented by its C^{α} carbon and an interaction between any two residues is represented by a spring between their C^{α} atoms. A spring between two residues is placed if they are within a certain cutoff distance. To model these modes of motions, there are several steps:

First, the potential energy V of the system, assumed to be lowest for the starting form, is assumed to be harmonic, and increases as a function of the square of the displacement:

$$V = \frac{\gamma}{2} DHD^T \qquad (1)$$

where the vector, D is the displacement, γ is the force constant for all the spring, and H is the Hessian matrix containing the second derivatives of the energy function. If the structure has n residues, the Hessian matrix H contains n x n super-elements; each element is of size 3 x 3. The (ij) th super element of the Hessian matrix can be derived from the following equation:

$$H_{ij} = \begin{bmatrix} \frac{\partial^2 V}{\partial X_i \partial X_j} & \frac{\partial^2 V}{\partial X_i \partial Y_j} & \frac{\partial^2 V}{\partial X_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Y_i \partial X_j} & \frac{\partial^2 V}{\partial Y_i \partial Y_j} & \frac{\partial^2 V}{\partial Y_i \partial Z_j} \\ \frac{\partial^2 V}{\partial Z_i \partial X_j} & \frac{\partial^2 V}{\partial Z_i \partial Y_j} & \frac{\partial^2 V}{\partial Z_i \partial Z_j} \end{bmatrix} \qquad (2)$$

where X_i , Y_i, and Z_i are the positional components of residue i; V represents the harmonic potential between residues i and j. Thus V can be expressed as follows:

$$V = \frac{\gamma}{2}\left(s_{ij} - s_{ij}^0\right)^2 \quad = \frac{\gamma}{2}\left(\left[\left(X_j - X_i\right)^2 + \left(Y_j - Y_i\right)^2 + \left(Z_j - Z_i\right)^2\right]^{\frac{1}{2}} - s_{ij}^0\right)^2 \qquad (3)$$

where s^0_{ij} is the equilibrium distance between residues i and j. Therefore, H can be decomposed as follows:

$$H = M\Lambda M^{-1} \qquad (4)$$

where Λ is a diagonal matrix of the eigenvalues and the columns of M are the eigenvectors. Each eigenvector represents one mode of motions of the structure except that the first 6 modes represent the rigid body translations and rotations of the structure. Thus the fluctuation of the structure is expressed as a set of 3n-6 modes formed by the 7^{th} to the $(3n)^{th}$ eigenvector – each eigenvector giving the direction and magnitude of the corresponding mode. The eigenvalues are sorted in descending order and each represents the importance and frequency of the corresponding mode.

Cutoff Distance Selection for ANM Model: *E.coli* FBA (PDB Id 1B57) and ***E.coli* TBA (PDB Id 1RVG)** – By comparing different ENM fluctuations with experimental B-factors, we find that either 13 or 14 Å are equally appropriate cutoff values. However, the cutoff value 14 Å gives a better distinction between monomeric and dimeric fluctuations. For ***T.aquaticus* FBA (PDB Id 1GVF)** a cutoff value of 15 Å yields the best prediction for the experimental B-factors; this also yields better discrimination among structures.

Authors' contributions

ARK and RLJ both contributed to the design, execution and writing of this work.

Bibliography

[1] J. J. Marsh and H. G. Lebherz, "Fructose-bisphosphate aldolases: an evolutionary history," *Trends Biochem. Sci*, vol. 17, no. 3, pp. 110-113, Mar. 1992.

[2] W. J. RUTTER, "EVOLUTION OF ALDOLASE," *Fed. Proc.*, vol. 23, pp. 1248-1257, Nov. 1964.

[3] D. E. Morse and B. L. Horecker, "The mechanism of action of aldolases," *Adv. Enzymol. Relat Areas Mol. Biol.*, vol. 31, pp. 125-181, 1968.

[4] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics.*, vol. 23, no. 21, pp. 2947-2948, Nov. 2007.

[5] S. A. Benner, M. A. Cohen, and G. H. Gonnet, "Amino acid substitution during functionally constrained divergent evolution of protein sequences," *Protein Eng*, vol. 7, no. 11, pp. 1323-1332, Nov. 1994.

[6] S. M. Zgiby, G. J. Thomson, S. Qamar, and A. Berry, "Exploring substrate binding and discrimination in fructose1, 6-bisphosphate and tagatose 1,6-bisphosphate aldolases," *Eur. J Biochem.*, vol. 267, no. 6, pp. 1858-1868, Mar. 2000.

[7] S. Qamar, K. Marsh, and A. Berry, "Identification of arginine 331 as an important active site residue in the class II fructose-1,6-bisphosphate aldolase of Escherichia coli," *Protein Sci*, vol. 5, no. 1, pp. 154-161, Jan. 1996.

[8] D. R. Hall, G. A. Leonard, C. D. Reed, C. I. Watt, A. Berry, and W. N. Hunter, "The crystal structure of Escherichia coli class II fructose-1, 6-bisphosphate aldolase in complex with phosphoglycolohydroxamate reveals details of mechanism and specificity," *J Mol. Biol.*, vol. 287, no. 2, pp. 383-394, Mar. 1999.

[9] S. Zgiby, A. R. Plater, M. A. Bates, G. J. Thomson, and A. Berry, "A functional role for a flexible loop containing Glu182 in the class II fructose-1,6-bisphosphate aldolase from Escherichia coli," *J Mol. Biol.*, vol. 315, no. 2, pp. 131-140, Jan. 2002.

[10] A. R. Plater, S. M. Zgiby, G. J. Thomson, S. Qamar, C. W. Wharton, and A. Berry, "Conserved residues in the mechanism of the E. coli Class II FBP-aldolase," *J Mol. Biol.*, vol. 285, no. 2, pp. 843-855, Jan. 1999.

[11] D. R. Hall, C. S. Bond, G. A. Leonard, C. I. Watt, A. Berry, and W. N. Hunter, "Structure of tagatose-1,6-bisphosphate aldolase. Insight into chiral discrimination, mechanism, and specificity of class II aldolases," *J Biol. Chem*, vol. 277, no. 24, pp. 22018-22024, Jun. 2002.

[12] T. Izard and J. Sygusch, "Induced fit movements and metal cofactor selectivity of class II aldolases: structure of Thermus aquaticus fructose-1,6-bisphosphate aldolase," *J Biol. Chem*, vol. 279, no. 12, pp. 11825-11833, Mar. 2004.

[13] A. Fiser, R. K. Do, and A. Sali, "Modeling of loops in protein structures," *Protein Sci.*, vol. 9, no. 9, pp. 1753-1773, Sept. 2000.

[14] N. Eswar, D. Eramian, B. Webb, M. Y. Shen, and A. Sali, "Protein structure modeling with MODELLER," *Methods Mol. Biol.*, vol. 426, pp. 145-159, 2008.

[15] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints," *J. Mol. Biol.*, vol. 234, no. 3, pp. 779-815, Dec. 1993.

[16] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, Jan. 2000.

[17] S. J. Cooper, G. A. Leonard, S. M. McSweeney, A. W. Thompson, J. H. Naismith, S. Qamar, A. Plater, A. Berry, and W. N. Hunter, "The crystal structure of a class II fructose-1,6-bisphosphate aldolase shows a novel binuclear metal-binding active site embedded in a familiar fold," *Structure.*, vol. 4, no. 11, pp. 1303-1315, Nov. 1996.

[18] "The PyMOL Molecular Graphics System, Version 1.4, Schrödinger, LLC.," 2012.

[19] "tagatose 1,6-bisphosphate in pubchem," 2013.

[20] "fructose 1,6-bisphosphate entry in pubchem," 2013.

[21] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC. Bioinformatics.*, vol. 9, p. 40, 2008.

[22] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nat. Protoc.*, vol. 5, no. 4, pp. 725-738, Apr. 2010.

[23] S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative TASSER simulations," *BMC. Biol.*, vol. 5, p. 17, 2007.

[24] Y. Zhang, A. Kolinski, and J. Skolnick, "TOUCHSTONE II: a new approach to ab initio protein structure prediction," *Biophys. J*, vol. 85, no. 2, pp. 1145-1164, Aug. 2003.

[25] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophys. J.*, vol. 80, no. 1, pp. 505-515, Jan. 2001.

CHAPTER 3. TRIOSEPHOSPHATE ISOMERASE STRUCTURE SPACE DIVERSITY: OLIGOMERIZATION, DYNAMICS, AND FUNCTIONALITY – AN EVOLUTIONARY PERSPECTIVE

Manuscript prepared for submission to a peer reviewed scientific journal

Ataur R. Katebi and Robert L. Jernigan

Abstract

Triosephosphate isomerase (TIM) catalyzes the reaction to convert dihydroxyacetone phosphate (DHAP) into glyceraldehyde 3-phosphate (GAP) and vice versa. In most organisms, its functional oligomeric state is a homodimer; however, tetramer formation in hyperthermomophiles is required for its functional activity. A tetrameric TIM structure in these organisms also provides added stability to the structure, enabling it to function at extreme temperatures. The protein data bank (PDB) has many experimental structures for dimeric TIMs and a few structures for tetrameric TIM structures. Also there are a substantial number of engineered monomeric TIM structures that have been determined. Engineered monomeric TIM structures from T.brucei mesophile are found to retain the residual catalytic activity. We applied Prinipal Component Analysis to find that the TIM structure space clearly gets divided into two groups – open TIM structures and closed TIM structures. The distribution of the structures in the closed set is much denser than that in the open set. We also apply ENM to four different TIM structures – an engineered monomeric structure (monoTIM), a dimeric structure (TbTIM) from a mesophile – T.brucei, and two tetrameric TIM structures (TmTIM and PwTIM) from distinct hyperthermophiles – T.maritima and P.woesei, respectively. We find that dimerization not only stabilizes the TIM structures, it also enhances their functional dynamics. Moreover,

tetramerization of the hyperthermophilic TIM structures increases their functional loops dynamics, enabling them to function in the destabilizing environment of extreme temperatures. Computations also show that the functional loops are highly coordinated in the TIM structures. Together with the high coordination of the TIM functional loops, stabilized dimeric and tetrameric TIM structures achieve their high production rates in proportion to the increased functional loop dynamics.

Key Words: triosephosphate isomerase; dihydroxy acetone phosphate; glyceraldehyde 3-phosphate; proton shuttling.

Abbreviations:

ATP	adenosine triphosphate	PDB	Protein Data Bank
DHAP	dihydroxy acetone phosphate	PGH	phosphoglycolohydroxamic acid
ENM	Elastic Network Model	PwTIM	*P. woesei* TIM – PDB Id 1HG3
GAP	glyceraldehyde 3-phosphate	TbTIM	*T. brucei* TIM – PDB Id 1TPE
monoTIM	Engineered monomeric TIM	TIM	triosephosphate isomerase
PCA	Principal Component Analysis	TmTIM	*T. maritima* TIM – PDB Id 1B9B

3.1 Introduction

3.1.1 Diversity of TIM Sequence and Structure Space

Triosephosphate isomerase (TIM) is the fifth enzyme in the eukaryotic glycolysis pathway which consists of 10 sequential steps that convert one molecule of glucose into two molecules of pyruvate. In the process it uses two ATP molecules and produces four ATP molecules with a net gain of two ATP molecules. TIM isomerizes dihydroxy acetone phosphate (DHAP) into glyceraldehyde 3-phosphate (GAP). TIM is an essential enzyme in most organisms and many organisms maintain a defense mechanism against the destruction of this enzyme by including in their genomes duplicate genes for this enzyme. Duplicate activity may also be needed to

maintain the required level of TIM activity in an organism [3]. In all organisms, TIM is found to be in the dimeric state as an active enzyme except in the hyperthermophilic organisms such as *Thermotoga maritima* [4], *Pyrococcus woesei* [5], *Thermoproteus tenax* [6], and *Methanocaldococcus jannaschii* [7], where its functional state is a tetramer.

Panel A of Fig. 1 shows the distribution of the number of TIM sequences of different lengths found in Pfam database and panel B shows such a distribution for the number of TIM structures in the PDB database. Out of 2,285 valid TIM sequences in Pfam with the length ranging from 222 to 276, the most frequent length is 251 with the frequency of 238. Other high frequency TIM lengths are 248 (156), 249 (160), 250 (162), 252 (145), 253 (145), 254 (116), 255 (110), and 256 (123). On the other hand, the highest frequency PDB lengths are 248 (26), 250 (24), 247 (19), and 238 (12). From these two distributions shown on panel A and panel B, it is interesting to notice that the PDB database has a representative TIM structure dataset from its Pfam TIM sequence dataset. Panel C of this figure shows the principal component analysis of 267 subunits of the TIM PDB structures. The first and the second PCs (PC1 and PC2) divide the structures into open and closed sets. The 198 subunits of the open set show their more diverse nature than the 69 subunits of the closed set. The span of RMSD distribution for these two sets also shows that the span of the closed subunits is much tighter than the span of open subunits – 0 ~ 0.69 Å (closed set) and 0 ~ 2.31 Å (open set).

3.1.2 Monomeric, Dimeric, and Tetrameric TIM Architectures

The structure of a TIM subunit follows the (α/β)-barrel architecture, with the two types of secondary structures alterating along the sequence. It has a central barrel consisting of eight β-strands surrounded by eight helices α1 ~ α8. Figure 2A shows such an arrangement of a TIM

subunit. The C-termini of the strands make the front of the barrel and the other ends of the strands constitute the back of the barrel.

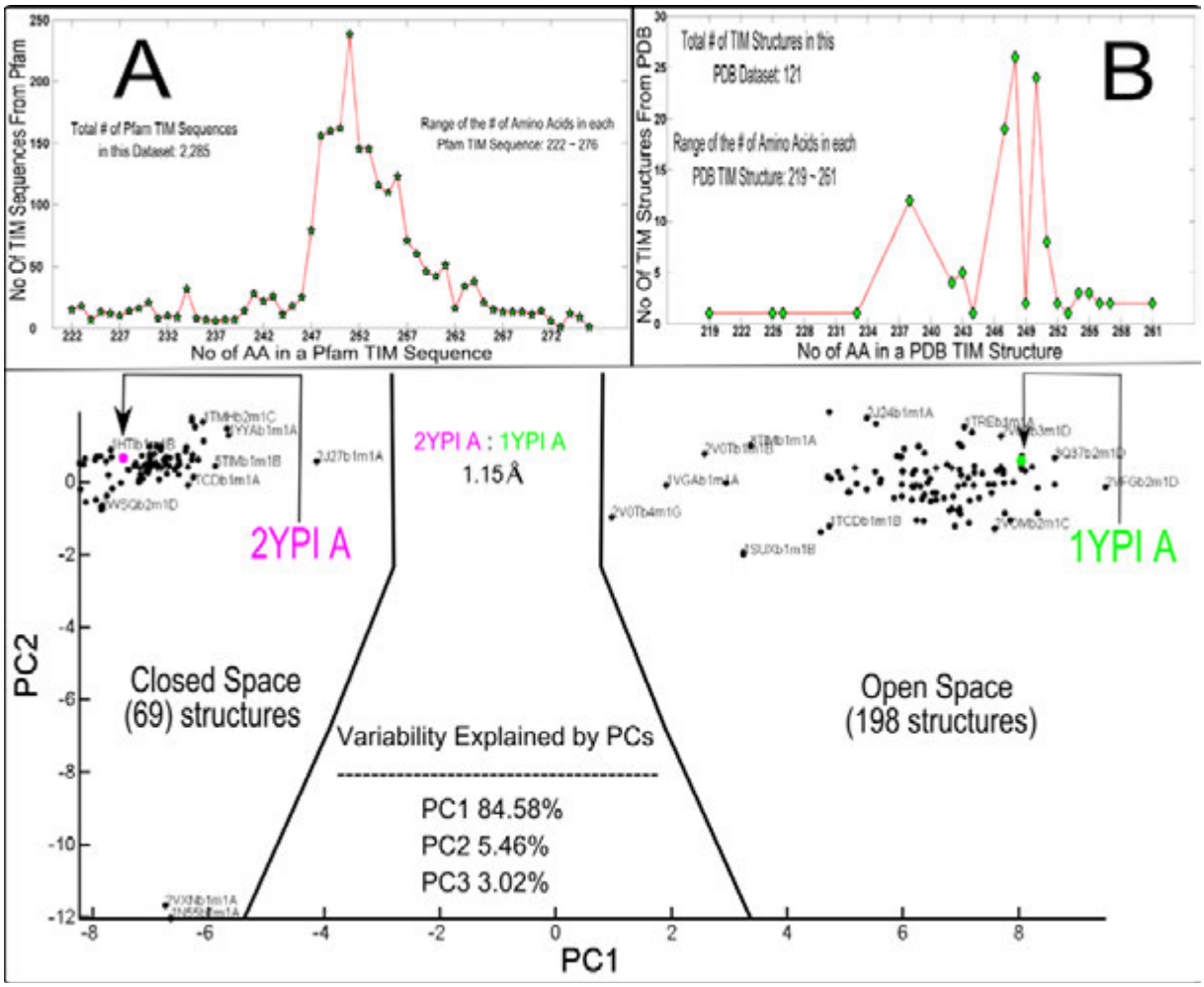


Figure 16. Distributions of TIM sequences and TIM structures (A) Length distributions of Pfam TIM sequences; (B) Length distributions of PDB TIM structures; (C) Principal component analysis of TIM PDB structures. This shows clearly that the closed structures are more similar to one another than are the set of open structures. The distribution also indicates that the primary coordinate for the transition is along PC1.

There are eight loops at the front – front loops FL1 ~ FL8 and eight loops at the back – back loops BL1 ~ BL8 Each front loop runs from a strand to a helix and each back loop runs from a helix to a strand. Thus the whole structure has such an arrangement of the strands, loops, and helices: N terminus – (β1 – FL1 – α1) – BL1 – (β2 – FL2 – α2) – BL2 – (β3 – FL3 – α3) – BL3 – (β4 – FL4 – α4) – BL4 – (β5 – FL5 – α5) – BL5 – (β6 – FL6 – α6) – BL6 – (β7 – FL7 – α7) –

BL7 – (β8 – FL8 – α8) – C terminus. The number of amino acids that constitute these secondary structure segments varies somewhat from one TIM structure to another. But the overall architecture of the structure is strictly conserved. Table I shows the positions of the secondary structure segments in the sequences for three different organisms that are considered in this research – *T.brucei* TIM, *T.maritima* TIM, and *P.woesei* TIM. The front loops are grouped into two sets – the loops forming the interface (front loops 1, 2, 3, and 4) and the loops that drive the catalysis (front loops 6, 7, and 8).

In mesophilic organisms, functional TIM enzyme is a homo dimer. However TIM is found to be an active homo tetrameric structure in some extremophilic organisms. Dimerization of TIM occurs through the association of two TIM monomers by a type 1 interface. Two type 1 dimeric TIM structures bind together by interactions along the two type 2 interfaces to form a homo tetrameric structure. The locations of these interfaces are marked in panels B, C, and D, of Fig. 2. The loops shown are all front loops, and this designation is dropped hereafter.

Figure 3 shows these two interfaces in greater detail. Four interface loops (1, 2, 3, and 4) from each subunit take part in forming the type 1 interface for the subunit-subunit association that is present in the dimer. Loop 3 from one subunit docks between loop 1 and loop 4 of the other subunit and loop 2 gets buried between them. Figures 3A and 3B show such an association. In the tetrameric organization, there are two type 1 dimeric TIM structures bind together via two type 2 interfaces. A type 2 interface is formed by the association of the C-terminus of loop 4, the N-terminus of helix 4, and helix 5 of one subunit with the same set of the interacting subunit. Figures 3A and 3C show the details of this construction.

3.1.3 Conserved Functional Mechanism across Species

The function and its mechanism of TIM structures are conserved across species. The three principal components of this phenomenon are:

Substrate trapping in the hydrophobic cage and product release by the concerted motions of functional loops 6 and 7.

Substrate specificity facilitated by loop 8.

Catalysis by substrate - proton transfer from DHAP to GAP and vice versa.

Roles of open and closed conformations in this mechanism.

(i) Substrate trapping in the hydrophobic cage and the product release by the concerted motions of functional loop 6 and loop 7.

The rate constant of the opening and closing motion of the active site loop 6 nearly matches with the production rate for TIM catalysis. This loop motion is coordinated with substrate binding, catalytic onset, and product release [9;10]. Crystallographic studies have shown that loop 6 and loop 7 have a closed conformation in the presence of ligand in the catalytic cavity and an open conformation in the absence of ligand [11;12]. Figures 4B and 8 show the catalytic pocket of the superimposed structures of the open and closed conformations. However, experiments also show that for the substrate to be trapped (bound) in the catalytic pocket, loop 6 closing is not necessary although the closed conformation is required for substrate catalysis (PDB Ids 1LYX, 1LZO [13]).

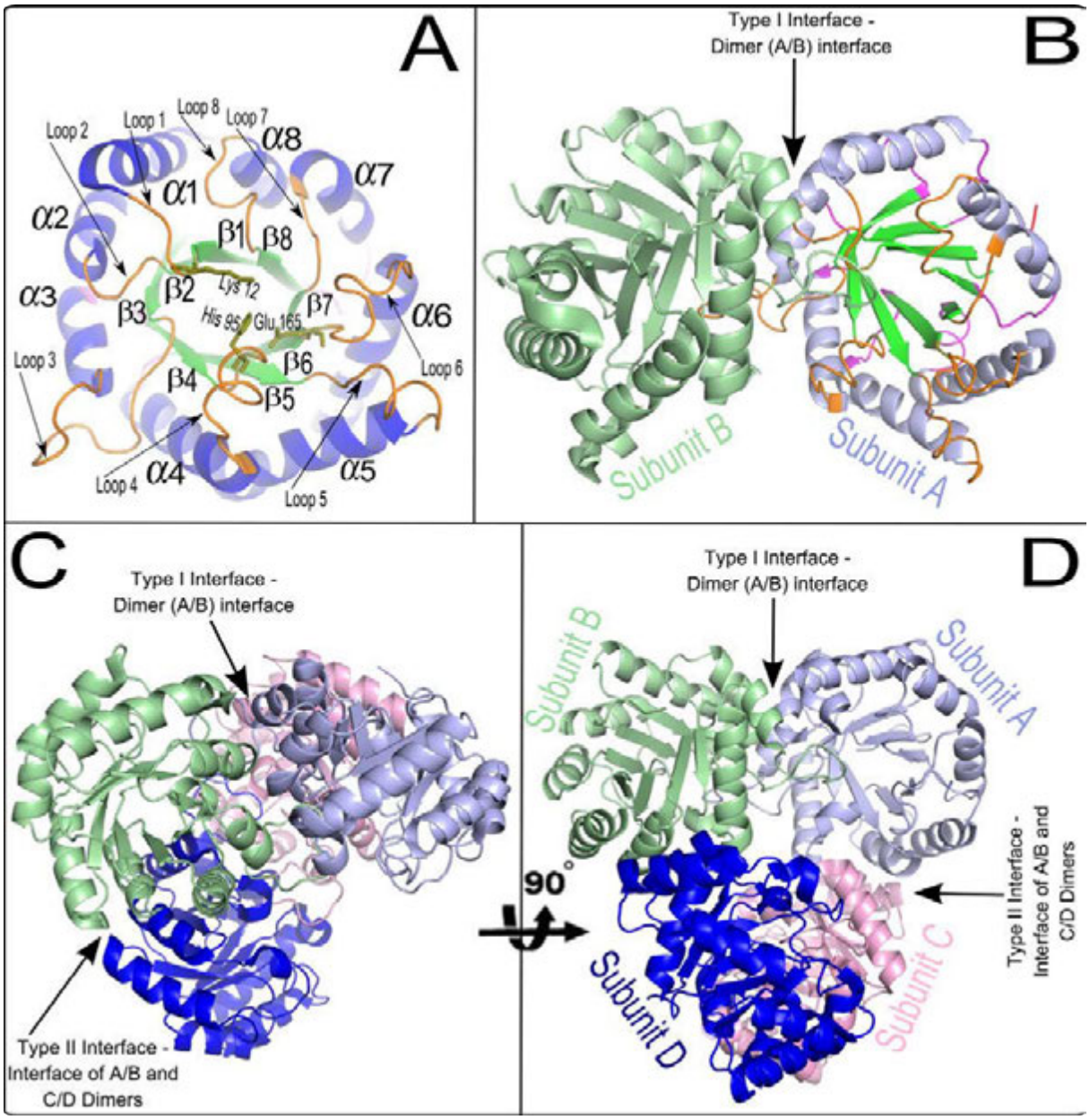


Figure 17. (A) Different structural components of a monomeric TIM subunit (based on *S.cerevisiae* TIM with PDB Id 1YPI). Eight strands β1 ~ β8 form the central barrel. The helices α1 ~ α8 surround the barrel. Front loops are labeled as Loop 1 ~ Loop 8. The catalytic residues are labeled Lys 12, His 95, and Glu 165. (B) Dimeric TIM architecture (base on *S.cerevisiae* TIM with PDB Id – 7TIM). A dimer is formed by the interactions along the type 1 interface of two subunits. (C) Tetrameric TIM structure (based on *P.woesei* TIM with PDB Id – 1HG3 [5]). Two Type 1 dimers form a tetrameric structure by the interaction along the two type 2 interface regions. (D) The tetrameric TIM structure in panel C after a 90° rotation around the indicated axis.

Table I. Positions of the Secondary Structure Segments in the Sequences of T.brucei TIM (1TPE), T.maritima TIM (1B9B) and P.woesei TIM (1HG3)						
Type of Secondary Structure	**Residue Indices in the Protein Sequence**					
	***T.brucei* TIM**		***T.maritima* TIM**		***P.woesei* TIM**	
	Indices	Length	Indices	Length	Indices	Length
N-terminal coil	1 – 7	7	1 – 5	5	1 – 7	7
Helices						
Helix 1	19 – 29	11	18 – 30	13	22 – 36	15
Helix 2	48 – 54	7	47 – 55	9	49 – 56	8
Helix 3	80 – 86	7	81 – 85	5	81 – 86	6
Helix 4	106 – 118	13	107 – 119	13	103 – 116	4
Helix 5	139 – 150	12	140 – 152	13	128 – 134	7
Helix 6	180 – 197	18	181 – 198	18	161 – 172	12
Helix 7	219 – 223	5	217 – 221	5	188 – 195	8
Helix 8	240 – 247	8	241 – 250	10	214 – 223	10
Back Loops						
Loop 1	30 – 37	8	31 – 37	7	37 – 40	4
Loop 2	55 – 59	5	56 – 60	5	57 – 60	4
Loop 3	87 – 89	3	86 – 90	5	87 – 90	4
Loop 4	119 – 121	3	120 – 122	3	117 – 118	2
Loop 5	151 – 161	11	153 – 163	11	135 – 139	5
Loop 6	198 – 207	10	199 – 207	9	173 – 177	5
Loop 7	224 – 229	6	222 – 230	9	196 – 200	5
Strands						
Strand 1	8 – 11	4	6 – 10	5	8 – 12	5
Strand 2	38 – 42	5	38 – 42	5	41 – 45	5
Strand 3	60 – 64	5	61 – 64	4	61 – 64	4
Strand 4	90 – 93	4	91 – 94	4	91 – 94	4
Strand 5	122 – 127	6	123 – 128	6	119 – 124	6
Strand 6	162 – 166	5	164 – 167	4	140 – 143	4
Strand 7	208 – 210	3	208 – 212	5	178 – 182	5
Strand 8	230 – 233	4	231 – 234	4	201 – 204	4
Front Loops						
Loop 1	12 – 18	7	11 – 17	7	13 – 21	9
Loop 2	43 – 47	5	43 – 46	4	46 – 48	3
Loop 3	65 – 79	15	65 – 80	16	65 – 80	16
Loop 4	94 – 105	12	95 – 106	12	95 – 102	8
Loop 5	128 – 138	11	129 – 139	11	125 – 127	3
Loop 6	167 – 179	13	168 – 180	13	144 – 160	17
Loop 7	211 – 218	8	213 – 216	4	183 – 187	5
Loop 8	234 – 239	6	235 – 240	6	205 – 213	9
C-terminal coil	248 – 250	3	251 – 255	5	224 – 225	2

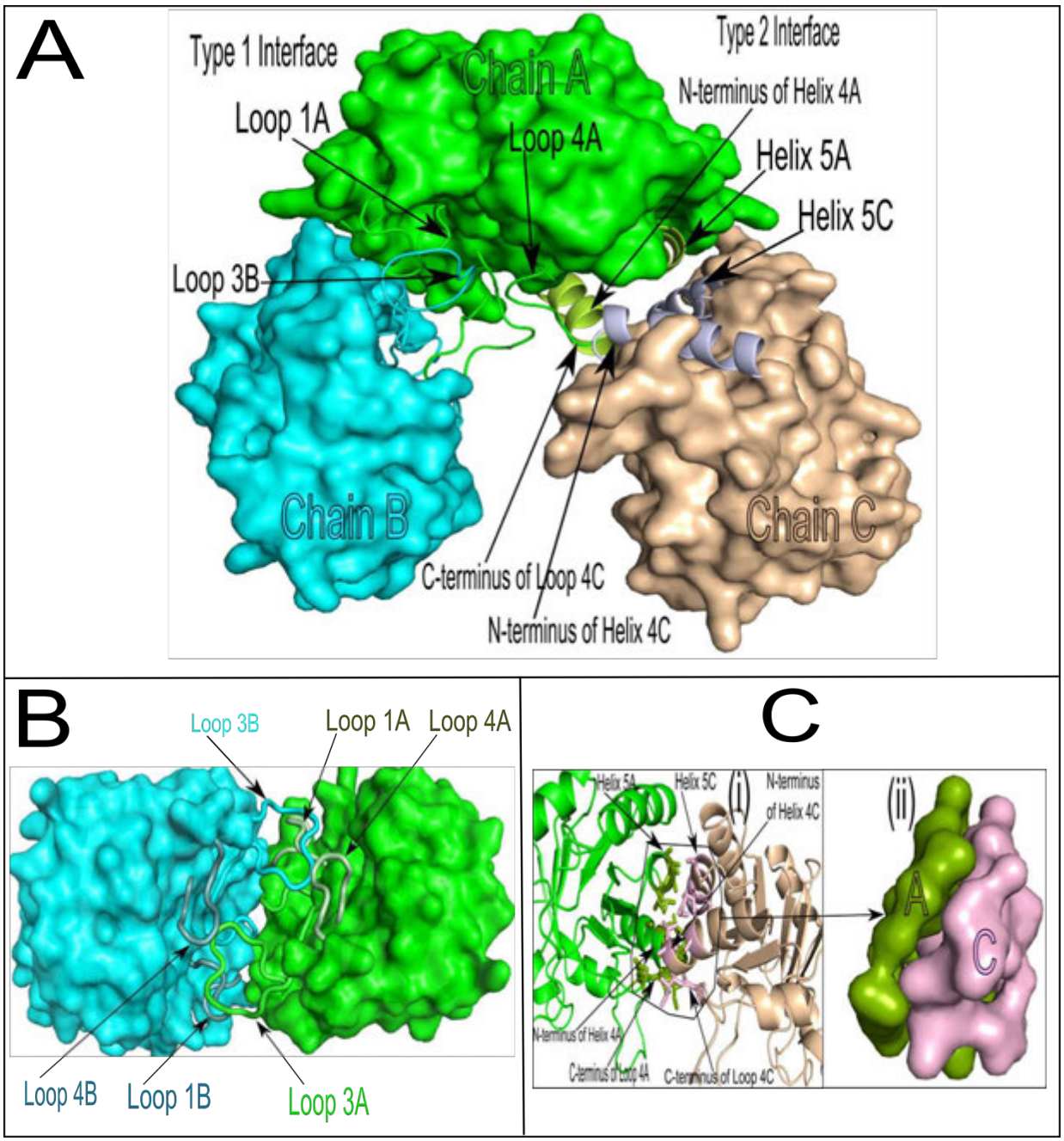


Figure 18. Structures at Subunit Interfaces. (A): Arrangement of type 1 and type 2 interfaces in a tetrameric TIM structure; (B) Interdigitation of loops in type 1 interface formation – Loop 3 of one subunit docks between loop 1 and loop 4 of the partner subunit; (C): (i) Structural components of the type 2 interface – Helices 5A and 5B, N-termini of helices 4A and 4C, C-termini of loop 4A and loop 4C construct the type 2 interface. (ii) Surface view of the type 2 interface shown in (i).

Experiments show that perturbation of dimerization of TIM structures reduces the rate of reaction of this enzyme by a factor of 1000 times. Dimerization of TIM enhances the motions in the loop 6 and loop 7 regions. It also increases the rigidity of loop 1, loop 4, and loop 8. This rigidity is needed to stabilize the position of the catalytic Lys on loop 1, and the catalytic His on loop 4, as well as Leu on loop 8 for the catalytic mechanism to function [14]. The active site loop dynamics is not only important for ligand release, they also limit the turnover rate of the protein [15].

Also the closing of loop 6 (an excursion of 7Å) stabilizes the reaction intermediate enediolates. This stabilization is facilitated by the tip of this loop which has a conserved 'phosphate gripper' motif -AXGXGKXA-[16]. This motif has some similarity to the consensus turn that interacts with phosphate groups in some kinases, many dehydrogenases, ras p21, and other nucleotide binding proteins [17-22].

(ii) Substrate specificity facilitated by loop 8

While the dynamics of loops 6 and loop 7 appear to directly determine the catalytic activity and rate, highly conserved loop 8 residues help to form a tight binding pocket for the phosphate moiety of the substrate. The fully conserved, solvent exposed Leu 238 (TbTIM residue indexing) of loop 8 limits the substrate binding specificity of TIM to only DHAP and GAP [23] (related PDB Id 1DKW).

The following residues from loops 6, 7, and 8 form H-bonds with the phosphate oxygen of the substrate in the closed conformation of the TIM structure: Gly 173 on loop 6; Gly 212 and Ser 213 on loop 7; and Gly 234 and Gly 235 on loop 8 [23].

(iii) Catalysis of the substrate – proton transfer from substrate DHAP to catalytic product GAP

Substrate catalysis in the catalytic pocket has two components. Physicochemical structure of the catalytic cavity: proper positioning of catalytic residues required for the proton transfer to happen [24]. The concerted motions of loop 6, loop 7, and loop 8 [25].

The catalytic residues Lys on loop 1, His on loop 4, and Glu on loop 6 complete the proton transfer by their coordination. These residues of the catalytic pocket are shown in Figs. 4A and 4B.

The concerted motions of loops 6 and 7 are important mainly for two reasons. First, the conformational flexibility of the catalytic Glu on loop 6 and its concerted motion with loop 7 facilitate proton shuttling [26]. Second, the concerted motion of loop 7 and the 'phosphate gripper' on loop 6 synchronizes the substrate trapping with the catalytic activity [16].

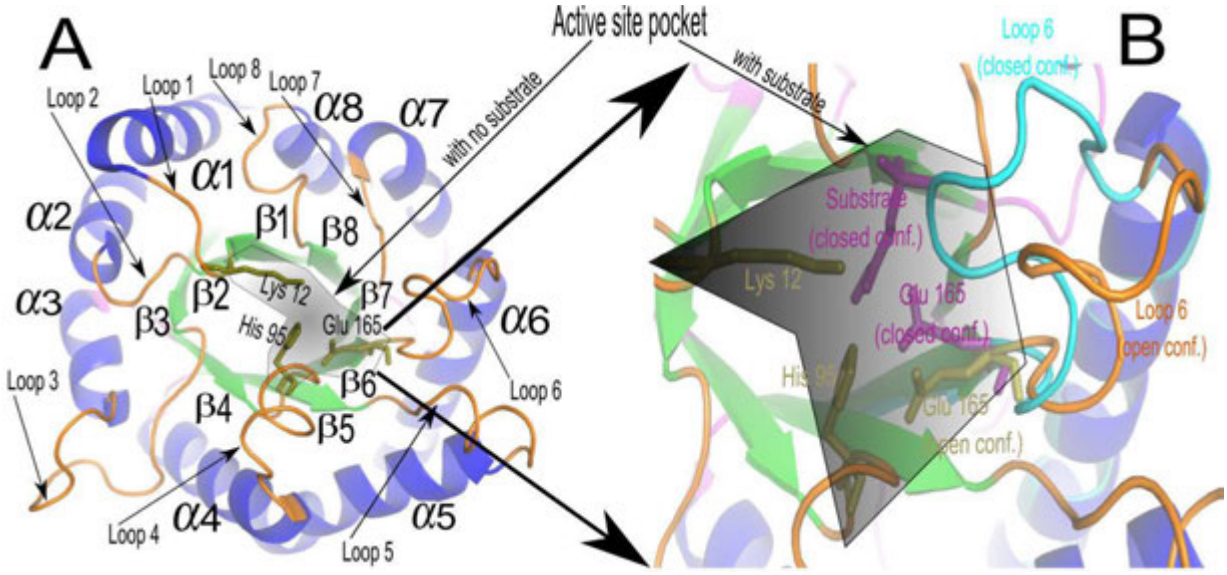


Figure 19. Structural details of active site of TIM/ (A) Location of the active site for its three catalytic residues in a TIM subunit (B) Active site pocket is shown in enlarged view with loop 6 in both open (orange) and closed (cyan) conformations. The structure and residue indexing is based on *S.cerevisiae* TIM structure PDB 1YPI.

(iv) Roles of open and closed conformations in the functional mechanism

TIM has two distinct conformations – open and closed. In the open conformation, loop 6 is wide open and appears in a much floppier state than it is in the closed conformation as shown in

Figs. 4(B) and 8. This flexibility of loop 6 is conducive to hunting and using the phosphate gripper to bring the substrate into the vicinity. Once the substrate is trapped in the cavity, loop 6 assumes the closed conformation and its 'phosphate gripper' keeps the substrate in place with the coordination of Leu 238 (TbTIM indexing) from loop 8. The closing of loop 6 affects the mechanism in two ways. First, it places the catalytic residue Glu 167 against to the substrate at the proper distance. Second, the correlated motions of loops 6 and 7 facilitate the proton transfer mechanism n in a coupled manner.

3.1.4 Relation between Protein Motions and their Functions

The architecture of a protein is responsible for its motions from the large scale domain movements to the local fluctuations. The way two or more domains attain their comparative movements is largely determined by the structure at the interfaces between the domains [27;28].The domain motions are important for the activities of a protein: its catalysis, the regulation of its activity, transport of metabolites, and forming protein assemblages.

In this research, we apply Elastic Network Models to investigate the dynamics of four structures in different oligomeric assemblies: monoTIM, TbTIM monomer and dimer; TmTIM monomer, dimer and tetramer; PwTIM monomer, dimer, and tetramer. We measure the average fluctuations of motions for different parts of the structure focusing particularly on (1) the parts important for interface formation (front loops 1, 2, 3, and 4) and (2) the region that is important (front loops 6, 7, and 8) for catalysis. We then compare these results for different monomeric states and measure and compare the correlations and overlaps of the motions of the different functional loops. From these computational results, we learn how oligomerization stabilizes the structures and also helps the structure to attain its native functional dynamics.

3.2 Results

3.2.1 Oligomerization and Stability across the Interface Region

Front loops 1, 2, 3, and 4

Type 1 interface is formed by the interdigitation of front loop 3 with front loops 1 and 4 of the partner subunit as shown in Fig. 5A. Panel B of the figure shows how this forms a locked situation between two subunits. Front loop 2 is buried by the front loop 3 from the partner subunit. Two such symmetric arrangements make a strong interface between the subunits to form a dimer. This dimerization locks these loops in place and reduces the dynamics of these loops. Panels C, D, and E of this figure show that dimerization decreases the fluctuations of front loop 1. Tetramerization of the structure reduces this dynamics further. This stabilization helps stabilize the catalytic residue Lys (K 13 in TbTIM, K 12 in TmTIM, and K 14 in PwTIM) on this loop. Panels F, G, and H show that fluctuations of loop 4, especially the C-terminus fluctuation, is reduced. ENM captures the decrease of motions in functional loop 1 and loop 4 with oligomerization. This decrease in motion stabilizes the catalytic residues Lys on loop 1 and His on loop 4 which is required for catalysis [9;10].

The tetrameric structure has almost the same stability of the catalytic Lys as in type 1 dimeric structure. Panel A shows that the catalytic Lys loses its required rigidity in engineered monomeric TIM monoTIM and monomer from TbTIM subunit. This could be a contributing factor to the reduced catalytic activity of monoTIM.

Panels A, B, and C of Fig. 6 show the reduced fluctuation of loop 2 after dimerization. Panels D, E, and F of Fig. 6 show the lowered fluctuations of loop 3 from the partner subunit after dimerization. Front loop 3 is the longest loop and has the highest mobility in the

standalone subunit. Dimerization lowers the fluctuations of the loops 1, 2, and 4 of the partner subunit. This is a common type of mechanism for transferring entropy from one region to another in a bound structure in comparison with an unbound structure. Because of the lower fluctuations of 1, 2, and 4 of the partner subunit, the catalytic loop region of the structure develops a functionally important increased fluctuation, which will be explained in the next section.

3.2.2 Oligomerization and Functional Loop Motions along the Catalytic Loops

Front loops 6, 7, and 8

Front loops 6, 7, and 8 surround the catalytic pocket. Figure 7A shows the functional loops in the open conformation of the structure (*S.cerevisie* PDB Id 1YPI). Here the substrate is copied from the closed TIM structure of the same organism (PDB Id 7TIM) whose catalytic site is shown in Fig. 7B. Functional loop 6 closes over the catalytic pocket in the closed conformation. The 'phosphate gripper' forms the tip of this loop. This consists of the following residues: 169 – AIGTGLAA –176. From the open to closed conformation, this region makes a large excursion towards the catalytic cavity – G 173 making the largest movement of 8.0Å and this motionis for the residues on either side in the loop are reduced as the distance along the sequence increases, as shown in Fig. 8. This conformational change of the phosphate gripper is important. In the open conformation, this region is disordered and may be used for substrate recruitment. Once the substrate is placed in the pocket, this loop stays in closed conformation by covering the opening of the catalytic pocket and thus protecting the catalytic mechanism from water and other molecules.

Panel C of Fig. 7 shows that monoTIM has much lower flexibility across the 'phosphate gripper' region (171:178 – AIGTGKVA) of loop 6 than for the TbTIM monomeric and dimeric

structures. TbTIM is an open TIM structure. Panels D and E show that this fluctuation for different oligomeric states of TmTIM and PwTIM is much lower. TmTIM and PwTIM are both closed structures. It has been experimentally shown that loop 6 flexibility is essential for substrate recruitment [9;10]. Therefore, there are two important conclusions: (1) engineered monoTIM loses most of its substrate recruitment capability because of the reduction in its loop flexibility, which occurs as a consequence of preventing its dimerization by shortening its interface loop 3. This also means that the shortening of the interface loop reduces has the broader effect of also reducing loop 6 flexibility and consequently reducing its catalytic activity. (2) Closed structures have a reduced loop 6 fluctuation compared to open structures. It means that higher fluctuation in the open state enable the structure to reach further out in its vicinity to recruit substrate. In the closed structure, the residual fluctuations of loop 6 gives enough dynamics to effect catalysis.

Panels D, E, and F of Fig. 7 show that the average ENM fluctuations of loop 7 in each case of TbTIM, TmTIM, and PwTIM has increased over dimerization and tetramerization,. Moreover, it is noticeable that engineered monoTIM has similar fluctuation as that of a monomer of TbTIM though it is reduced compared to the dimeric counterpart. Also, the loop 7 fluctuations in the closed structures (panels E and F) is much lower than that for the open structure (panel D). It has been found from experiments that loop 7 synchronizes its motions with the two hinge regions of loop 6 to drive the dynamics of loop 6 whose motion is important for substrate trapping, catalysis, and product release [9;10]. Therefore, there are several noteworthy points : (1) monoTIM still maintains loop 7 fluctuations, which thus retains its capacity for driving loop 6 dynamics in a reduced substrate catalysis. (2) Reduced fluctuations of this loop in the closed

structures (panels G and H) is sufficient to facilitate the catalytic mechanism and the required coordination with loop 6 that we will discuss in a following section.

Panel I compares the fluctuations of front loop 8 in an engineered monoTIM, monomer of TbTIM, and TbTIM dimer. It is clear that the fluctuations in this loop do not change much for these three cases. In case of TmTIM (panel J) and PwTIM (panel K), it is noticeable that the tetrameric structure has higher fluctuations than the monomer or the dimer of the same structure. Experiments show that the highly conserved loop 8 helps the TIM structure maintain a tight binding catalytic pocket. Especially, Leu (residue index 238 in TbTIM and 239 in TmTIM) helps maintin the high substrate specificity [14]. However, in case of PwTIM this is replaced by Lys (K 210). Higher fluctuations of loop 8 may cause the Leu to come out of its buried position to make room for the substrate to get properly positioned within the pocket.

From this we can conclude that tetramerization in hyperthermophilic organisms is required not only for structural stability but also for functional viability for survival in the thermally noisy extreme environment where those organisms have adapted.

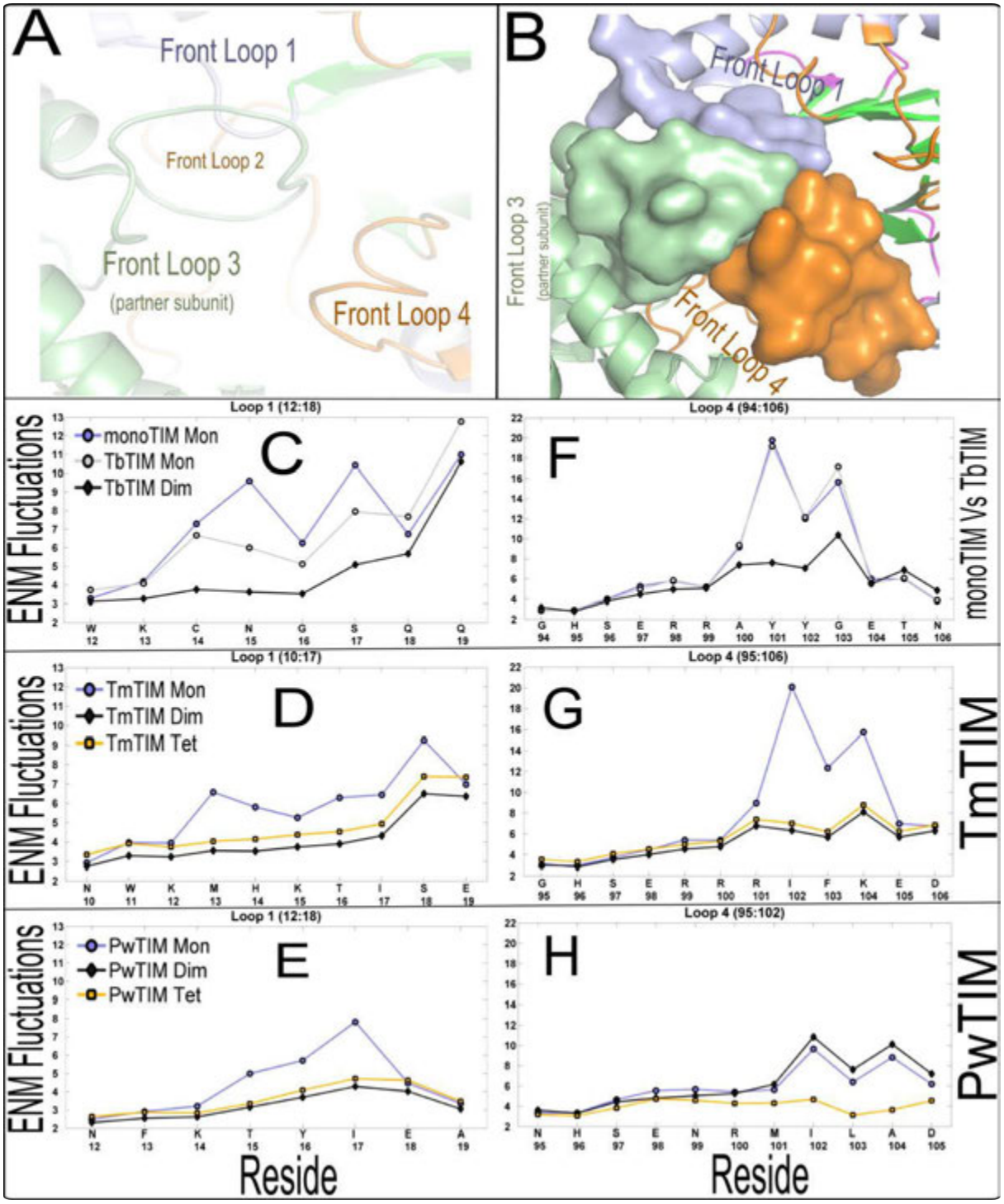


Figure 20. Dimeric interface loop fluctuations. (A) Cartoon view of interdigitation of front loops 1 and 4 with the front loop 3 from the partner subunit; (B) Surface view shows the docking of front loop 3 between the ridges of front loop 1 and 4 of the partner subunit. (C, D, E, F, G, & H). Fluctuations of the two front loops 1 and 4 are shown in different oligomeric states in three different organisms and an engineered monomer.

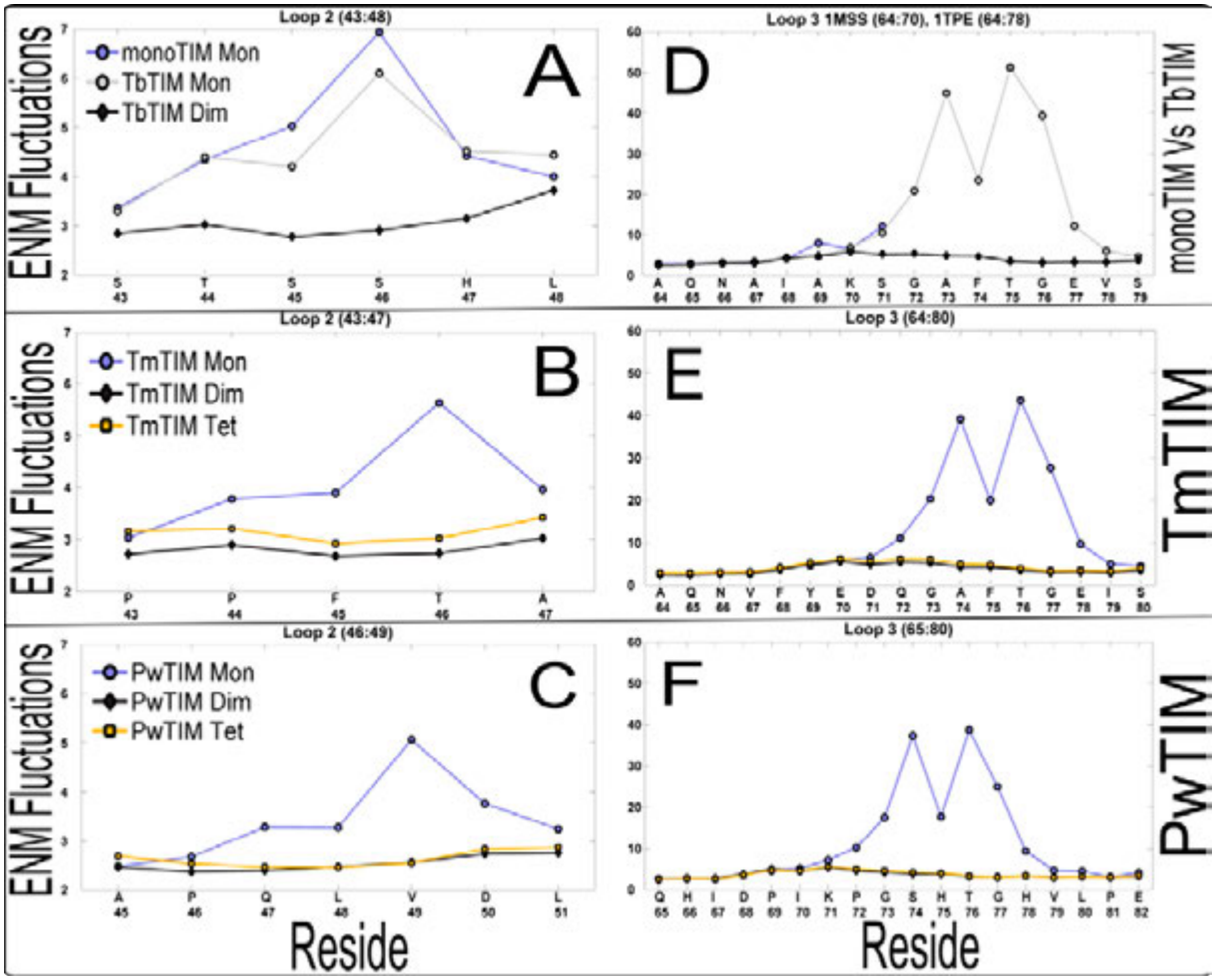


Figure 21. Changes in fluctuations of interface loops 2 and 8 in different oligomeric states.

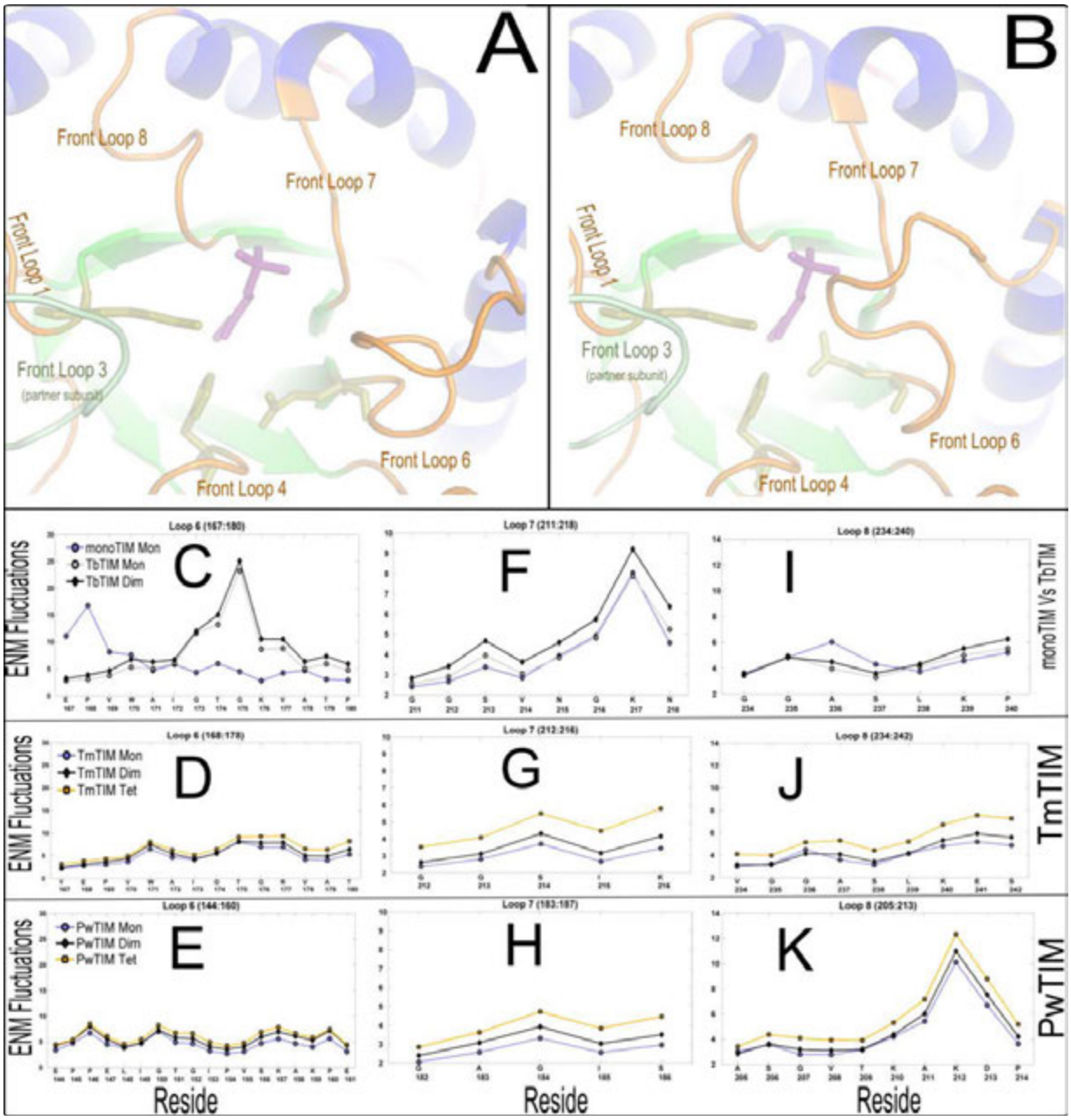


Figure 22. Change of fluctuations of functional loops. (A) Arrangement of functional loops 6, 7, and 8 around the catalytic pocket – open conformation (based on *S.cerevisiae* open TIM structure PDB 1YPI), substrate is inserted at the catalytic site by superimposing the open and closed conformations; (B) The same arrangement in closed conformation (based on *S.cerevisiae* TIM structure with PDB Id); (C, D, E) Change of fluctuations of functional loop 6; (F, G, H) Change of fluctuations of functional loop 7; (I, J, K) Change of fluctuations of functional loop 8.

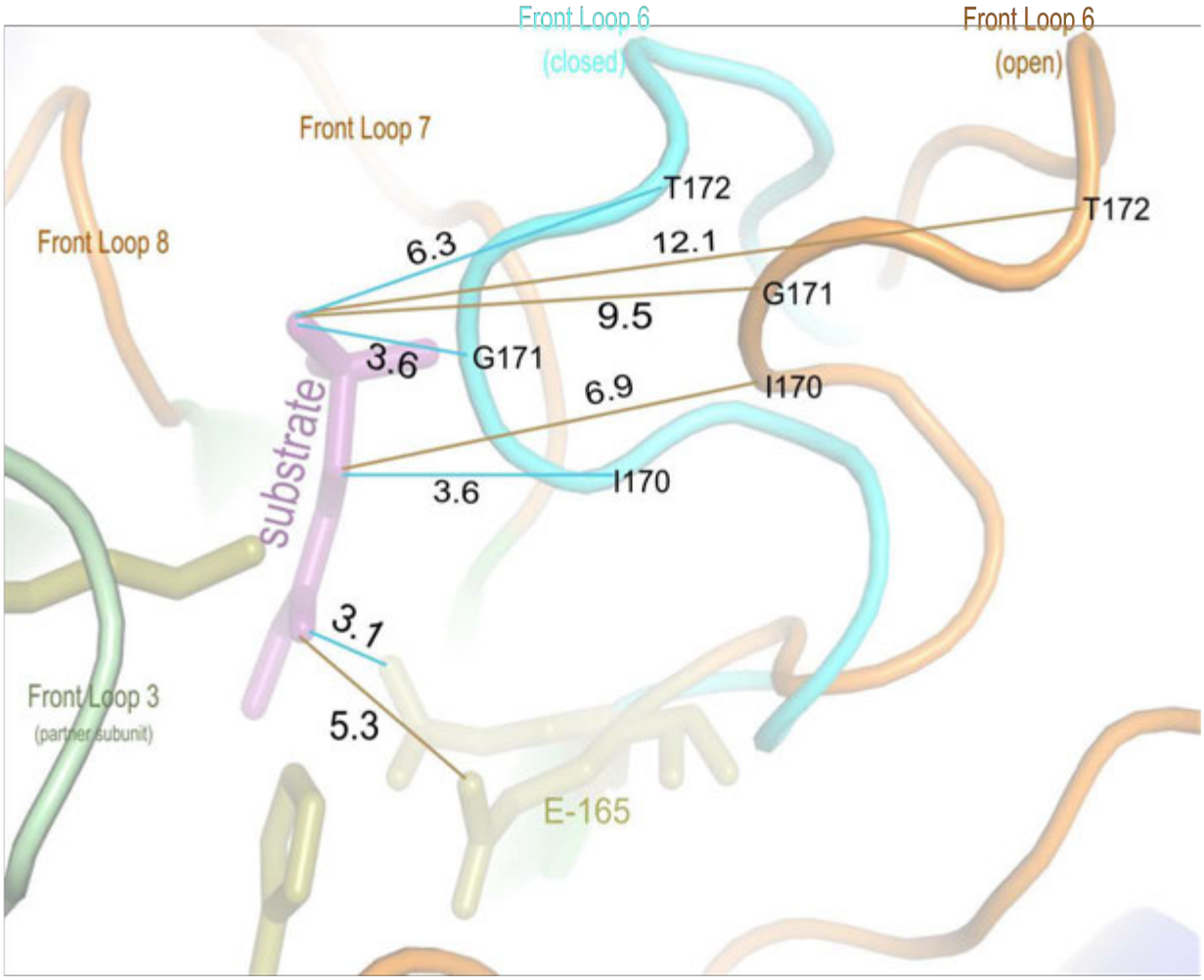


Figure 23. Change of distance between substrate and different residues of the 'phosphate gripper' between the open (orange) and closed (cyan) conformations. The residue indexing and open and closed conformations of the loops were generated using PyMol and *S.cerevisiae* PDB structure 1YPI (open conformation) and 7TIM (closed conformation). The excursions of the residues themselves in the 'phosphate gripper' towards the catalytic pocket are: A169 3.0Å, I170 4.2Å, G171 6.7Å, T172 6.7Å, G173 8.0Å, L174 5.0Å, A175 4.3Å, and A176 1.6Å. Colors for lines: orange – the distance between substrate and the residue from open loop, cyan – the distance between the substrate and residues from closed loop.

3.2.3 Concerted Motions of Functional Loops

Correlations between loop 6 and loop 7 dynamics (Table II)

Different experiments have shown that loop 7 plays a crucial role in the concerted motions of the N and C-terminal hinge residues of catalytic loop 6 essential to maintain the high efficiency of TIM production [31]. Our analysis of TIM dynamics by ENM detects the parts of loop 6 and loop 7 where their motions are highly correlated and how they are maintained for different oligomerization states.

Table II. Comparing correlations of ENM fluctuations of loop 6 and loop 7 of chain A within different oligomeric states of four TIM structures

Loop 6 Segments	Loop 7		
Loop 6 Definition (167:179)	monoTIM Monomer (210:218)		
167:175	0.6811		
168:176	0.9441		
Loop 6 Definition (167:179)	TbTIM		
	Monomer (210:218)	Dimer (210:218)	
167:175	0.7028	0.7425	
168:176	0.9240	0.9347	
Loop 6 Definition (167:180)	TmTIM		
	Monomer (212:217)	Dimer (212:217)	Tetramer (212:217)
167:172	0.5309	0.6093	0.7290
169:174	0.7680	0.5409	0.5252
170:175	0.7202	0.5724	0.6441
172:177	0.3291	0.4615	0.6359
Loop 6 Definition (145:161)	PwTIM		
	Monomer (182:186)	Dimer (182:186)	Tetramer (182:186)
145:149	0.7820	0.6603	0.6136
149:153	0.8663	0.9179	0.9266
155:159	0.5839	0.5946	0.7104
156:160	0.7918	0.8197	0.7023

monoTIM and TbTIM

168:176 – PVWAIGTGK – Loop 7 has the highest correlation with this segment of loop 6 in all three cases – monoTIM (0.9441), TbTIM monomer (0.9240) and TbTIM dimer (0.9347). This segment consists of the N-terminal hinge (168:170 – PVW) and the rigid tip (171:175 – AIGTG) of loop 6. Moreover, this segment includes the 'phosphate gripper' (169:176 – VWAIGTGK) of loop 6 ('phosphate gripper' motif – -AXGXGKXA-[16]).

Though monoTIM has reduced flexibility in both loop 6 and loop 7 as shown in Fig. 7, it still maintains a high correlation between these two loops. This indicates that each of all these three oligomeric states maintains high correlation between loop 6 and loop 7 that is required for the proton transfer mechanism of the TIM catalytic activity to operate.

Therefore the reduction of monoTIM activity may develop for other reasons:

Reduced substrate recruitment capability because of lowered flexibility of loop 6

Reduced rigidity of loop 1 and loop 4 as shown in panel C of Fig. 5 where the stability of catalytic residues Lys and His is necessary for their required proximity to the substrate.

TmTIM monomer, dimer, and tetramer

In TmTIM tetrameric structure, loop 7 shows high correlation with loop 6 in four segments:

167:172 – YEPVWA (0.7290) – This section of loop 6 has the catalytic residue Glu 168 and N-terminal hinge region (169:171 – PVW).

169:174 – PVWAIG (0.5252) – This segment contains the N-terminal hinge region (169:171 – PVW) and the most portion (172:174 –AIG) of the rigid tip (172:176 – AIGTG).

170:175 – VWAIGT (0.6441) – This segment contains the most portion (172:175 – AIGT) of the rigid tip (172:176 –AIGTG) and the N-terminal region (172:175 – AIGT) of the ‘phosphate gripper’ (172:179 – AIGTGRVA).

172:177 – AIGTGR (0.6359) – This segment contains the rigid tip (172:176 –AIGTG) of loop 6. Also, it contains N-terminus (172:177 – AIGTGR) of the ‘phosphate gripper’ (172:179 – AIGTGRVA).

In sum:

Interestingly, the segment of loop 6 with the catalytic residue Glu 168 achieves the highest correlation in tetrameric TmTIM.

The N-terminal region, the tip of loop 6, and the ‘phosphate gripper’ region of loop 6 also achieve high correlation with loop 7.

These regions also maintain high correlation in monomeric and dimeric structures.

PwTIM monomer, dimer, and tetramer

In PwTIM tetrameric structure, loop 7 shows high correlation with loop 6 in four segments:

145:149 – PPELI (0.6139) – This segment contains the catalytic residue Glu 147 and the two residues (148:149 – LI) of the N-terminal hinge (148:150 – LIG).

149:153 – IGTGI (0.9266) – This segment has the highest correlation. It contains the ‘rigid tip’ (148:151 – AIGTG) of the loop and the major portion (148:153 – LIGTGI) of the ‘phosphate gripper’ (148:155 – LIGTGIPV).

155:159 – VSKAK (0.7104) – This segment contains the C-terminal hinge region (156:158 – SKA) of loop 6.

156:160 – SKAKP (0.7023) – This segment contains the C-terminal hinge (156:158 – SKA) and a portion (159:160 – KP) of the coil region at the C-terminal end of the loop.

In sum:

Four regions of loop 6 that achieve high correlation with loop 7 are: (a) segment that has the catalytic residue Glu that initiates the proton transfer mechanism of the catalytic process, (b) the 'rigid tip' of loop 6 that is responsible to keep the catalytic pocket safe from invasion of water molecules, (c) segment that contains the 'phosphate gripper' that is responsible to keep the loop 6 bound facilitating the safety of the cavity from water invasion and high TIM turnout, and (d) the C-terminal end of loop 6.

These segments also achieve high correlation in monomeric and dimeric structures.

Correlations between loop 6 and loop 8 dynamics (Table III)

Loop 8 shows significant correlations in the following regions of loop 6:

monoTIM and TbTIM

167:179 – EPVWAIGTGKVAT – This segment of loop 6 achieves the highest correlation (0.8167) with loop 8 in dimeric TbTIM compared to monoTIM (0.4203) and monomeric TbTIM (0.7253). This segment contains whole loop 6 region and the catalytic residue Glu.

TmTIM monomer, dimer, and tetramer

168:176 – EPVWAIGTG – This region has the catalytic residue Glu 168, the N-terminal hinge region (169:171 – PVW), the 'rigid tip' (172:176 –AIGTG), and the N-terminus (172:176 – AIGTG) of the 'phosphate gripper' (172:179 – AIGTGRVA). Tetrameric

TmTIM achieves the highest correlation (0.8305) in this segment compared to monomeric (0.6931), and type1 (0.7677) and type 2 (0.7305) dimeric TmTIM structures.

169:177 – PVWAIGTGR – This segment has the N-terminal hinge region (169:171 – PVW), the 'rigid tip' (172:176 –AIGTG), and the N-terminus (172:177 – AIGTGR) of the 'phosphate gripper' (172:179 – AIGTGRVA). This region maintains high correlation in all four oligomeric states of TmTIM – 0.9517, 0.9090, 0.8921, and 0.9401, respectively.

PwTIM monomer, dimer, and tetramer

146:156 – PELIGTGIPVS – This segment has the negative correlation in each oligomeric state of PwTIM.

150:160 – GTGIPVSKAKP – This segment has low correlation in each oligomeric state of PwTIM.

Correlations between loop 7 and loop 8 dynamics (Table IV)

Loop 7 shows significant correlations with the following regions of loop 8.

monoTIM and TbTIM

Correlation between loop 7 and each comparable segment of loop 8 changes from positive in monoTIM to negative in monomeric and dimeric TbTIM .

232:240 – This segment of loop 8 achieves the most negative correlation in monomeric and dimeric TbTIM.

233:241 – This segment of loop 8 has the highest positive correlation (0.4642) in monoTIM.

TmTIM monomer, dimer, and tetramer

235:240 and 236:241 – In both loop 8 regions, tetrameric TmTIM achieves the highest correlations (0.9050 and 0.7174, respectively) compared to monomeric, type 1, and type 2 dimeric structures.

Table III. Comparing correlations of ENM fluctuations of loop 6 and loop 8 of chain A within different oligomeric states for four structures			
Loop 6 Segments	Loop 8		
Loop 6 Defintion (167:179)	monoTIM Monomer (232:242)		
167:177	0.4203		
Loop 6 Definition (167:179)	TbTIM		
	Monomer (232:242)	Dimer (232:242)	
167:177	0.7253	0.8167	
Loop 6 Definition (167:180)	TmTIM		
	Monomer (234:242)	Dimer (234:242)	Tetramer (234:242)
168:176	0.6931	0.7677	0.8305
169:177	0.9517	0.9090	0.9401
Loop 6 Definition (145:161)	PwTIM		
	Monomer (204:214)	Dimer (204:214)	Tetramer (204:214)
146:156	-0.5160	-0.4786	-0.4059
150:160	0.2580	0.4831	0.4413

PwTIM monomer, dimer, and tetramer

204:208 – This region of loop 8 maintains very high correlation in each PwTIM oligomeric state.

210:214 – This region maintains the second highest correlation in each PwTIM oligomeric state.

Surprisingly, correlations between loop 7 and 8 are low or negative in monoTIM and TbTIM (both monomeric and dimeric), respectively. On the other hand, this correlation is very high in TmTIM and PwTIM. This implies that correlations between loops 6 and 7, and loops 6 and 8, are stronger in TmTIM and PwTIM compared to monoTIM and TbTIM. The significance of this implication lies in the following proposition - higher oligomerization increases functional loop correlations in the tetrameric structure making it much more efficient than its dimeric counterpart.

Table IV. Comparing correlations of ENM fluctuations of loop 7 and loop 8 of chain A within different oligomeric states for four structures

Loop 8 Segments	Loop 7		
Loop 8 Definition (232:242)	monoTIM Monomer (210:218)		
232:240	0.3176		
233:241	0.4642		
Loop 8 Definition (232:242)	TbTIM		
	Monomer (203:211)	Dimer (203:211)	
232:240	-0.6216	-0.7058	
233:241	-0.3841	-0.6002	
Loop 8 Defintion (232:242)	TmTIM		
	Monomer (212:217)	Dimer (212:217)	Tetramer (212:217)
235:240	0.6558	0.8925	0.9050
236:241	0.2769	0.5950	0.7174
Loop 8 Definition (204:214)	PwTIM		
	Monomer (182:186)	Dimer (182:186)	Tetramer (182:186)
204:208	0.8217	0.9127	0.8980
210:214	0.6274	0.6546	0.5676

3.2.4 Overlap of Modes of ENM Motions

Our computation shows that the overlap of motions between chain A and chain B in a hypothetical type 1 dimeric structure is much higher than the overlap between chain A and chain C of a hypothetical type 2 dimer for *P.woesei* as shown in Fig. 9. However, in the tetrameric structure, overlaps between the chains across the barrel are the highest and they are almost symmetric for the two pairs – chain A compared with chain D and chain B compared with chain C in Fig. 9.

3.2.5 Changes of Loop Motions with Change in Correlations of Functional Loop Motions

The change in flexibilities of functional loops in oligomerization not only facilitates substrate binding but also increase the correlation between functional loop motions which is required for the synchrony of the catalytic mechanism. In tetrameric TIM, we can see a higher rigidity in loops 1, 4, and 8, with higher flexibility in loops 6 and 7. The correlations between loops 6, 7, and 8 also increase overall in the tetrameric TIM structures.

In hyperthermophilic TIM, oligomerization is required to achieve sufficient cohesion to carry out the catalysis. Two functionally inactive dimers come together to form a functionally active tetrameric complex through the type 2 interface. This tetramerization increases functional loop fluctuations and their correlations as well.

3.3 Discussion and Conclusion

3.3.1 Oligomerization Regulate Substrate Binding and Catalysis

Motions of loops 1, 2, 3, and 4: Dimerization of a TIM structure brings the catalytic residue Lys on loop 1 closer the catalytic pocket. Also, it reduces the fluctuations of the interface loops thus stabilizing the structure. Tetramerization stabilizes the structure further in hyperthermophilic organism. However, monoTIM interface loops are more mobile and thus could be a reason for its reduced catalytic activity. *We hypothesize conservation of entropy since the interaction energies will not change much.* In other words, if some flexibility is lost for some loops upon oligomerization then other parts become compensatingly more flexible.

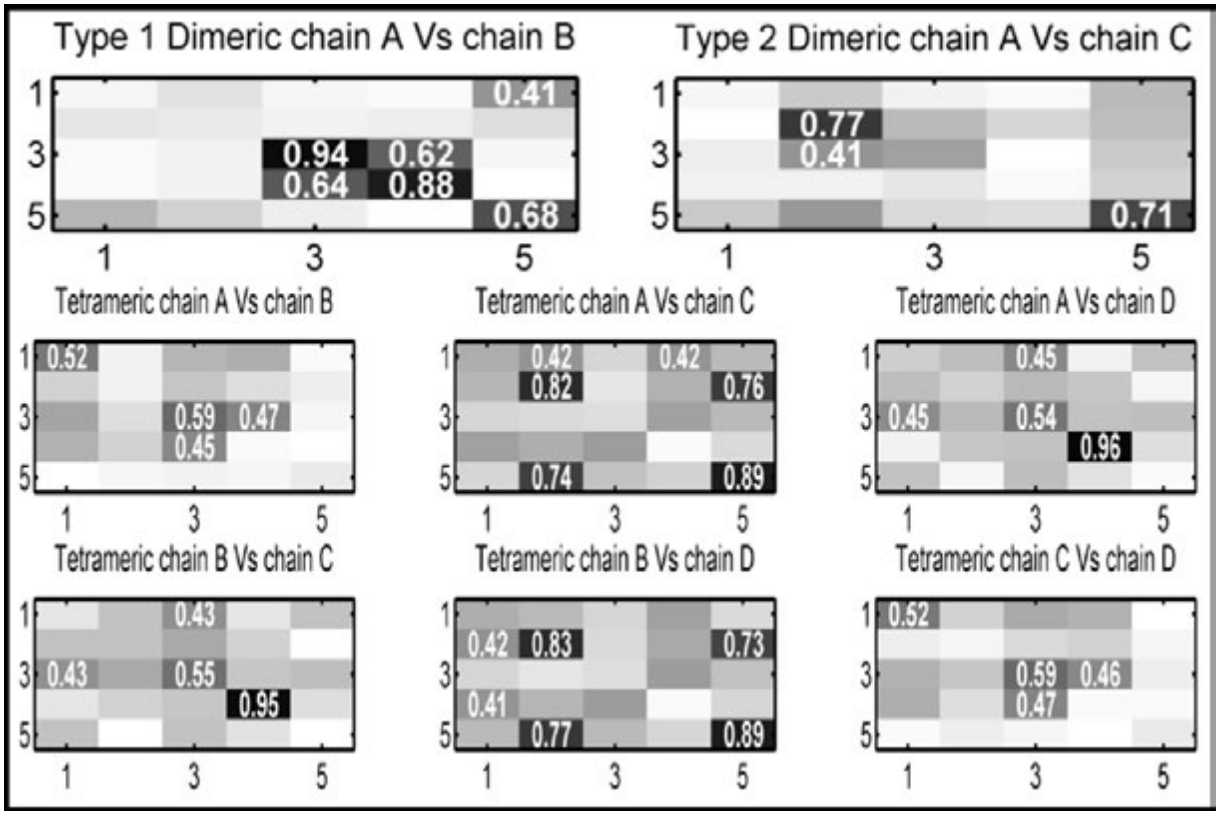


Figure 24. Overlaps of modes of ENM motions between chains in the PwTIM structure.

Motions of loops 6, 7 and 8: Motions of loops 6 and loop 7 depend on two events –

Oligomerization: Motions of these two loops increase from oligomerization which is necessary for substrate recruitment.

Substrate Binding: Motions of these two loops decrease during substrate binding which is necessary to protect the catalytic cavity from penetration by unwanted small molecules.

Leu on loop 8 that is responsible for the highly specific shape of the TIM catalytic cavity is believed to be stabilized because of the reduced loop 8 motions. However, ENM models show that oligomerization in fact may increase the loop 8 motion in some cases. Dimerization decreases this motion; Tetramerization increases the fluctuation slightly.

Correlation of loop 6 and loop 7: Loop 6 and loop 7 maintain a high correlation, regardless of the oligomeric state of TIM. Interestingly, in different oligomeric states, loop 7 changes its highest correlation value with different parts of loop 6. This might imply that although different oligomeric states have high correlation between loop 7 and different parts of loop 6, loop 7 of the functionally active oligomers (monoTIM, TbTIM dimer, TmTIM tetramer, PwTIM tetramer) achieve the expected overall high correlation in the required region of loop 6.

Catalytic competency of TIM hinges on two important things: (i) stability of loop 1 and loop 4 (catalytic residues Lys on loop 1 and His on loop 4); and (ii) flexibility and coordination of loop 6 and loop 7. Tetramerization increases both.

The loop motion and their correlations are the important rate determining factors for this enzyme. In hyperthermophilic organisms, the enzyme needs to be efficient by increasing the fluctuation of the loops. Also because of the extreme operating temperature, the stability of the enzyme needs to be increased. Tetramerization serves both purposes.

3.4 Materials and Methods

3.4.1 Data Set Preparation

We have prepared two datasets for this exploration – a TIM sequence database and a TIM structure database.

TIM Sequence Database: We have downloaded the TIM sequences from Pfam dated 10/19/2012 (http://pfam.sanger.ac.uk/) [32]. The number of sequences is 4,005 which also includes sequence fragments and putative sequences. After removing the sequence fragments and putative sequences, we have 2,285 full length TIM sequences whose lengths range from 222 to 276. However, most sequences fall between lengths of 247 and 257.

TIM Structure Database: We have downloaded 121 TIM structures from the Protein Data Bank (PDB) (www.pdb.org)[33] dated 10/14/2011. After extracting each chain from the structures we have a database of 307 individual TIM chains. This database contains monomeric chains (engineered or mutated so that dimerization does not happen), dimeric chains (either wild type or mutated), and tetrameric chains for four hyperthermophilic organisms – 1B9B (*Thermotoga maritima*), 1HG3 (*Pyrococcus woesei*), 1W0M (*Thermoproteus tenax*), and 2H6R (Methanocaldococcus jannaschii). Functional loop 6 is the most disordered region of the TIM enzyme and many structures are missing this loop,either in part or in its entirety. After removing all chains with missing loop 6, we have 297 chains remaining. If we remove the tetrameric chains and some very irregular monomeric chains, we have 263 chains. 105 PDB ids of the TIM structures that are used to extract these 263 chains are as follows:

1AG1 1AMK 1AW1 1AW2 1BTM 1CI1 1DKW 1HTI 1I45 1IF2 1IIG 1IIH 1KV5 1LYX 1LZO 1M6J 1M7O 1M7P 1ML1 1MO0 1MSS 1N55 1O5X 1QDS 1R2R 1R2S 1R2T 1SQ7

1SSD 1SSG 1SU5 1SUX 1SW0 1SW3 1SW7 1TCD 1TIM 1TMH 1TPB 1TPC 1TPD 1TPE 1TPF 1TPH 1TPU 1TPV 1TPW 1TRD 1TRE 1TSI 1TTI 1TTJ 1VGA 1WOA 1WOB 1YDV 1YPI 1YYA 2BTM 2DP3 2I9E 2J24 2J27 2JGQ 2JK2 2OMA 2V0T 2V2C 2V2D 2V2H 2V5B 2V5L 2VEK 2VEM 2VEN 2VFD 2VFE 2VFF 2VFG 2VFH 2VFI 2VOM 2VXN 2WSQ 2X1R 2X1S 2X1T 2X1U 2YPI 3GVG 3KRS 3M9Y 3PF3 3PVF 3PWA 3PY2 3Q37 3TH6 3TIM 3YPI 4TIM 5TIM 6TIM 7TIM 8TIM

We have normalized these chains by aligning each chain with the yeast PDB structure 2YPI chain A. This normalized dataset is available in the supplementary information. We call this dataset of 263 normalized chains the normal TIM. All structural analysis shown in Fig. 2 is based on this set of normal TIM chains.

PDB Structures to Measure the Allosteric Effect of Oligomerization: We have selected four PDB structures to observe the allosteric effect of TIM Oligomerization. 1MSS (Engineered monomer from *Trypanosoma brucei* TIM), 1TPE (*Trypanosoma brucei* TIM), 1B9B (*Thermotoga maritima*), and 1HG3 (*Pyrococcus woesi*).

1MSS (monoTIM): Each chain is an engineered monomeric TIM. This structure is in its open conformation with no substrate bound in the active site. The active site loops, loops 1 and 4, of this structure adopt very different conformations from the wild type TIM. Shortening of the length of loop 3 from 15 residues to 8 residues causes the disruption of the subunit-subunit contacts in monomTIM; consequently, the essential side chains of Lys 13 on loop 1 and His 95 on loop 4 move away from their catalytically active positions. However, these loops adopt the wild type conformations in the closed engineered monomer (1TTI and 1TTJ), and are very different from the monoTIM. This possibility to form the closed form may explain the residual catalytic activity of monoTIM. The optimal catalysis of the wild type TIM is facilitated by the

required rigidity of loop 1, loop 4, and loop 8, occurring because of the subunit-subunit contacts at the dimer interface [34].

1TPE (TbTIM): is a *T. brucei* TIM in its open conformation [35].

1B9B (TmTIM): No sequence preferences are correlated with thermal stability considering ten TIM structures ranging from psychrophiles to hyperthermophiles based on analysis of amino acid composition or the analysis of the loops and secondary structure elements. A common feature for both psychrophilic and hyperthermophilic TIM (in this case, *T.maritima* TIM) is the large number of salt bridges compared with the number found in mesophilic TIMs. Thermophilic TIMs have the highest amount of accessible hydrophobic surface buried during the folding and assembly process [4]. The N-terminus of hyperthermophilic *T.maritima* TIM has been shown to be covalently linked to the C-terminus of phosphoglycerate kinase (PGK), forming a bifunctional PGK-TIM fusion protein which is a tetramer consisting of four PGK-TIM chains [36]. The *T.maritima* TIM structure (1B9B) possess more salt bridges and more buried hydrophobicity upon both folding and assembly [4].

1HG3 (PwTIM): The extreme thermostability is achieved by the creation of a compact tetramer where two classical TIM dimers interact via an extensive hydrophobic interface. The tetramer is formed through largely hydrophobic interactions between some of the pruned helical regions. The equivalent helical regions in less thermostable dimeric TIMs represent regions of high average temperature factors [5].

3.4.2 Principal Component Analysis (PCA) - Exploring the TIM Structure Space by PCA

PCA – Principal Component Analysis

PCA is a multivariate technique to analyze a dataset where the observations are described quantitatively by a set of inter-correlated variables. The goals of PCA are to (i) extract the most important information from the data; (ii) remove noise and compress the size of the data set by keeping only this important information; (iii) simplify the description of the data set; and (iv) analyze the structure of the observations and the variables. This method generates a set of new orthogonal variables called principal components (PCs). Each PC is a linear combination of the original variables. Hence, PCA can be considered as a mapping of the data points from the original variable space to the PC space. PCs are computed in such a way that when each data point is projected on PC1, the resulting values form a new variable that has the maximum variance among all possible choices for the first axis. Similarly, when each data point is projected on PC2, the resulting values form another new variable that has the maximum variance among all possible choices for the second axis, and so forth. The number of PCs could be as many as the number of the original variables. However, usually a few PCs are sufficient to understand the internal structure of the data [37]. The mathematical derivation of the PCs is summarized in the supplementary materials.

Exploring the TIM Structure Space by PCA

The coordinates of 14 residues that span the loop 6 residues are extracted from each of the 263 chains of the TIM structure dataset, normalizedTIMchains. The motifs that are used to find the start of loop 6 are: EPIWAIG, EPVWSIG, EPPELIG, EPLWAIG, EPLWAIG, EPVWAAT, EPVWAIT, EPVWAVG, EPLFAIG, EAVWAIG, DPVWAIG.

Here, the 14 residue positions on loop 6 are the variables of the data set for PCA. 263 segments, each spanning the loop 6 of each structure, are data points. We use *princomp* function of Matlab Statistical Toolbox to compute the PCs (PC1, PC2, PC3, ….) [2010a, The MathWorks, Natick, MA].

Mathematical Derivation of Principal Components

The mathematical derivation of principal components is summarized below according to Kevin P. Murphy [38]. Suppose, we have a D-dimensional dataset. We want to do PCA to get a projection of this dataset onto an orthogonal basis of size $K < D$, such that we preserve as much information as possible. More precisely, PCA assumes that any given vector $x_i \varepsilon R^D$ can be approximated as a linear combination of K basis vectors $v_j \varepsilon R^D$ as follows:

$$\hat{x}_i = V z_i$$

where $z_i \varepsilon R^K$ is the low dimensional representation of $x_i \varepsilon R^D$. The goal in PCA is to find an orthogonal set of K linear basis vectors $v_j \varepsilon R^D$, and the corresponding scores $z_i \varepsilon R^K$, such that we minimize average construction error,

$$J(V, Z) = \frac{1}{N} \sum_{i=1}^{N} ||x_i - \hat{x}_i||^2$$

subject to the constraint that $v_i^T v_i = 1$ and $v_i^T v_j = 0$ if $i \neq j$. Let us start by considering the first principal component, $v_1 \varepsilon R^D$, and the corresponding projected points $z_1 \varepsilon R^N$ where N is the number of data points in the D-dimensional dataset (another way of saying, N is the number of observations). The reconstruction error is then given by

$$J(v_1, z_1) = \frac{1}{N}\sum_{i=1}^{N}\left|\left|x_i - z_{i1}v_1\right|\right|^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}[x_i^T\, x_i - \, 2z_{i1}v_1^T x_i \, + \, z_{i1}^2]$$

Since $v_1^T v_1 = 1$, taking derivatives w.r.t z_{i1} and equating to zero gives

$$\frac{\partial}{\partial_{z_{i1}}} J(v_1, z_1) = \, v_1^T x_i$$

So the optimal reconstruction weights are obtained by orthogonally projecting the data points onto the first principal direction, v_1. Plugging back in gives

$$J(v_1) = \, const - v_1^T C v_1$$

where $C = \frac{1}{N}X^T X$. Dropping the constant term and adding the constraint yields

$$\tilde{J}(v_1) = \, v_1^T C v_1 + \, \lambda_1(v_1^T v_1 - 1)$$

where λ_1 is the Lagrange multiplier. Taking derivative and equating to zero, we have

$$Cv_1 = \lambda_1 v_1$$

Hence, the direction that maximizes the variance is an eigenvector of the covariance matrix of the dataset and it is proved as the first principal component. In order to find the 2nd principal component v_2, we further minimize the reconstruction error, subject to $v_1 v_2 = 0$ and $v_2^T v_2 = 1$. The error turns out,

$$J(v_2, z_2) = \frac{1}{N}\sum_{i=1}^{N}\left||x_i - z_{i1}v_1 - z_{i2}v_2|\right|^2$$

Taking $\frac{\partial J}{\partial z_2} = 0$, results in $z_{i2} = v_2^T x_i$. We obtain the 2nd principal component by projecting on the 2nd principal direction. Resultant equation turns out:

$$J(v_2) = const - v_2^T C v_2$$

Throwing away the constant term and removing the constraints yields,

$$\tilde{J}(v_2) = v_2^T C v_2 + \lambda_2(v_2^T v_2 - 1) + \lambda_{12}(v_2^T v_1 - 0)$$

Hence, the 2nd principal component is given by the eigenvector with the 2nd largest eigenvalue:

$$Cv_2 = \lambda_2 v_2$$

Following the similar procedure, we can compute the other principal components as well.

3.4.3 Modeling Dynamics

Given two structures, we compute their normal modes using the Anisotropic Network Model (ANM) [39]. The normal modes for each structure are represented as a set of vectors. The normal modes from one ANM model can be compared to the normal modes from another ANM model using equations 1, 2, and 3, which describe the 'Overlap' between the directions of the i^{th} mode of one model and the j^{th} mode of another model, 'Cumulative Overlap' between the first k

normal modes of one model and the i^{th} mode of another model, and overlap between the space spanned by the first I normal modes of one model and the space spanned by the first J normal modes of another model (their 'Root Mean Square Inner Product, RMSIP'), respectively as described by Tama and Sanejouand [40] and Leo-Macias, et. al. [41].

$$O_{ij} = \frac{|\mathrm{M_i}.\mathrm{M_j}|}{\|\mathrm{M_i}\|\|\mathrm{M_j}\|} \quad (1)$$

$$CO(k) = \sqrt{\sum_{j=1}^{k} O_{ij}^2} \quad (2)$$

$$RMSIP(I,J) = \sqrt{\frac{1}{I}\sum_{i=1}^{I}\sum_{j=1}^{J}(M_i M_j)^2} \quad (3)$$

We adapted these metrics to compare the functional loop dynamics of two different but structurally similar proteins – FBA and TIM. We select the equal length matching segments of loops from a specific protein pair and extract the normal modes for only those segments from the ENM results. We compute Overlap, Cumulative Overlap, and RMSIP for each of the re-orthonormalized sets of modes.

Cutoff selected for ANM: 12Å

Authors' contributions

ARK and RLJ both contributed to the design, execution and writing of this work.

Bibliography

[1] Kurkcuoglu O, Jernigan RL, and Pemra D, "Collective Dynamics of Large Proteins from Mixed Coarse-Grained Elastic Network Model," *QSAR & Combinatorial Science*, vol. 24, pp. 443-448, Jun. 2005.

[2] O. Kurkcuoglu, R. L. Jernigan, and P. Doruker, "Loop motions of triosephosphate isomerase observed with elastic networks," *Biochemistry*, vol. 45, no. 4, pp. 1173-1182, Jan. 2006.

[3] E. E. Figueroa-Angulo, P. Estrella-Hernandez, H. Salgado-Lugo, A. Ochoa-Leyva, P. A. Gomez, S. S. Campos, G. Montero-Moran, J. Ortega-Lopez, G. Saab-Rincon, R. Arroyo, C. G. Benitez-Cardoza, and L. G. Brieba, "Cellular and biochemical characterization of two closely related triosephosphate isomerases from Trichomonas vaginalis," *Parasitology*, pp. 1-10, Aug. 2012.

[4] D. Maes, J. P. Zeelen, N. Thanki, N. Beaucamp, M. Alvarez, M. H. Thi, J. Backmann, J. A. Martial, L. Wyns, R. Jaenicke, and R. K. Wierenga, "The crystal structure of triosephosphate isomerase (TIM) from Thermotoga maritima: a comparative thermostability structural analysis of ten different TIM structures," *Proteins*, vol. 37, no. 3, pp. 441-453, Nov. 1999.

[5] H. Walden, G. S. Bell, R. J. Russell, B. Siebers, R. Hensel, and G. L. Taylor, "Tiny TIM: a small, tetrameric, hyperthermostable triosephosphate isomerase," *J. Mol. Biol.*, vol. 306, no. 4, pp. 745-757, Mar. 2001.

[6] H. Walden, G. L. Taylor, E. Lorentzen, E. Pohl, H. Lilie, A. Schramm, T. Knura, K. Stubbe, B. Tjaden, and R. Hensel, "Structure and function of a regulated archaeal triosephosphate isomerase adapted to high temperature," *J. Mol. Biol.*, vol. 342, no. 3, pp. 861-875, Sept. 2004.

[7] P. Gayathri, M. Banerjee, A. Vijayalakshmi, S. Azeez, H. Balaram, P. Balaram, and M. R. Murthy, "Structure of triosephosphate isomerase (TIM) from Methanocaldococcus jannaschii," *Acta Crystallogr. D. Biol. Crystallogr.*, vol. 63, no. Pt 2, pp. 206-220, Feb. 2007.

[8] R. C. Davenport, P. A. Bash, B. A. Seaton, M. Karplus, G. A. Petsko, and D. Ringe, "Structure of the triosephosphate isomerase-phosphoglycolohydroxamate complex: an analogue of the intermediate on the reaction pathway," *Biochemistry*, vol. 30, no. 24, pp. 5821-5826, Jun. 1991.

[9] S. Rozovsky, G. Jogl, L. Tong, and A. E. McDermott, "Solution-state NMR investigations of triosephosphate isomerase active site loop motion: ligand release in relation to active site loop dynamics," *J Mol. Biol.*, vol. 310, no. 1, pp. 271-280, Jun. 2001.

[10] S. Rozovsky and A. E. McDermott, "The time scale of the catalytic loop motion in triosephosphate isomerase," *J Mol. Biol.*, vol. 310, no. 1, pp. 259-270, Jun. 2001.

[11] D. Joseph, G. A. Petsko, and M. Karplus, "Anatomy of a conformational change: hinged "lid" motion of the triosephosphate isomerase loop," *Science*, vol. 249, no. 4975, pp. 1425-1428, Sept. 1990.

[12] M. E. Noble, J. P. Zeelen, and R. K. Wierenga, "Structures of the "open" and "closed" state of trypanosomal triosephosphate isomerase, as observed in a new crystal form: implications for the reaction mechanism," *Proteins*, vol. 16, no. 4, pp. 311-326, Aug. 1993.

[13] S. Parthasarathy, G. Ravindra, H. Balaram, P. Balaram, and M. R. Murthy, "Structure of the Plasmodium falciparum triosephosphate isomerase-phosphoglycolate complex in two crystal forms: characterization of catalytic loop open and closed conformations in the ligand-bound state," *Biochemistry*, vol. 41, no. 44, pp. 13178-13188, Nov. 2002.

[14] B. V. Norledge, A. M. Lambeir, R. A. Abagyan, A. Rottmann, A. M. Fernandez, V. V. Filimonov, M. G. Peter, and R. K. Wierenga, "Modeling, mutagenesis, and structural studies on the fully

conserved phosphate-binding loop (loop 8) of triosephosphate isomerase: toward a new substrate specificity," *Proteins*, vol. 42, no. 3, pp. 383-389, Feb. 2001.

[15] S. Rozovsky, G. Jogl, L. Tong, and A. E. McDermott, "Solution-state NMR investigations of triosephosphate isomerase active site loop motion: ligand release in relation to active site loop dynamics," *J Mol. Biol.*, vol. 310, no. 1, pp. 271-280, Jun. 2001.

[16] J. R. Knowles, "Enzyme catalysis: not different, just better," *Nature*, vol. 350, no. 6314, pp. 121-124, Mar. 1991.

[17] P. H. Seeburg, W. W. Colby, D. J. Capon, D. V. Goeddel, and A. D. Levinson, "Biological properties of human c-Ha-ras1 genes mutated at codon 12," *Nature*, vol. 312, no. 5989, pp. 71-75, Nov. 1984.

[18] J. Reinstein, I. R. Vetter, I. Schlichting, P. Rosch, A. Wittinghofer, and R. S. Goody, "Fluorescence and NMR investigations on the ligand binding properties of adenylate kinases," *Biochemistry*, vol. 29, no. 32, pp. 7440-7450, Aug. 1990.

[19] W. Moller and R. Amons, "Phosphate-binding sequences in nucleotide-binding proteins," *FEBS Lett.*, vol. 186, no. 1, pp. 1-7, Jul. 1985.

[20] R. K. Wierenga, P. Terpstra, and W. G. Hol, "Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint," *J Mol. Biol.*, vol. 187, no. 1, pp. 101-107, Jan. 1986.

[21] D. F. Lowry, M. R. Ahmadian, A. G. Redfield, and M. Sprinzl, "NMR study of the phosphate-binding loops of Thermus thermophilus elongation factor Tu," *Biochemistry*, vol. 31, no. 11, pp. 2977-2982, Mar. 1992.

[22] M. Saraste, P. R. Sibbald, and A. Wittinghofer, "The P-loop--a common motif in ATP- and GTP-binding proteins," *Trends Biochem. Sci*, vol. 15, no. 11, pp. 430-434, Nov. 1990.

[23] B. V. Norledge, A. M. Lambeir, R. A. Abagyan, A. Rottmann, A. M. Fernandez, V. V. Filimonov, M. G. Peter, and R. K. Wierenga, "Modeling, mutagenesis, and structural studies on the fully conserved phosphate-binding loop (loop 8) of triosephosphate isomerase: toward a new substrate specificity," *Proteins*, vol. 42, no. 3, pp. 383-389, Feb. 2001.

[24] G. Jogl, S. Rozovsky, A. E. McDermott, and L. Tong, "Optimal alignment for enzymatic proton transfer: structure of the Michaelis complex of triosephosphate isomerase at 1.2-A resolution," *Proc. Natl. Acad. Sci U. S. A*, vol. 100, no. 1, pp. 50-55, Jan. 2003.

[25] I. Kursula, M. Salin, J. Sun, B. V. Norledge, A. M. Haapalainen, N. S. Sampson, and R. K. Wierenga, "Understanding protein lids: structural analysis of active hinge mutants in triosephosphate isomerase," *Protein Eng Des Sel*, vol. 17, no. 4, pp. 375-382, Apr. 2004.

[26] R. Venkatesan, M. Alahuhta, P. M. Pihko, and R. K. Wierenga, "High resolution crystal structures of triosephosphate isomerase complexed with its suicide inhibitors: The conformational flexibility of the catalytic glutamate in its closed, liganded active site," *Protein Sci*, vol. 20, no. 8, pp. 1387-1397, Aug. 2011.

[27] M. Gerstein, A. M. Lesk, and C. Chothia, "Structural mechanisms for domain movements in proteins," *Biochemistry*, vol. 33, no. 22, pp. 6739-6749, Jun. 1994.

[28] M. Gerstein, B. F. Anderson, G. E. Norris, E. N. Baker, A. M. Lesk, and C. Chothia, "Domain closure in lactoferrin. Two hinges produce a see-saw motion between alternative close-packed interfaces," *J. Mol. Biol.*, vol. 234, no. 2, pp. 357-372, Nov. 1993.

[29] J. Janin and C. Chothia, "The structure of protein-protein recognition sites," *J. Biol. Chem.*, vol. 265, no. 27, pp. 16027-16030, Sept. 1990.

[30] C. L. Lo, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J. Mol. Biol.*, vol. 285, no. 5, pp. 2177-2198, Feb. 1999.

[31] Y. Wang, R. B. Berlow, and J. P. Loria, "Role of loop-loop interactions in coordinating motions and enzymatic function in triosephosphate isomerase," *Biochemistry*, vol. 48, no. 21, pp. 4548-4556, Jun. 2009.

[32] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy, "The Pfam protein families database," *Nucleic Acids Res.*, vol. 32, no. Database issue, p. D138-D141, Jan. 2004.

[33] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, Jan. 2000.

[34] T. V. Borchert, K. V. Kishan, J. P. Zeelen, W. Schliebs, N. Thanki, R. Abagyan, R. Jaenicke, and R. K. Wierenga, "Three new crystal structures of point mutation variants of monoTIM: conformational flexibility of loop-1, loop-4 and loop-8," *Structure.*, vol. 3, no. 7, pp. 669-679, Jul. 1995.

[35] K. V. Kishan, J. P. Zeelen, M. E. Noble, T. V. Borchert, and R. K. Wierenga, "Comparison of the structures and the crystal contacts of trypanosomal triosephosphate isomerase in four different crystal forms," *Protein Sci*, vol. 3, no. 5, pp. 779-787, May 1994.

[36] H. Schurig, N. Beaucamp, R. Ostendorp, R. Jaenicke, E. Adler, and J. R. Knowles, "Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium Thermotoga maritima form a covalent bifunctional enzyme complex," *EMBO J.*, vol. 14, no. 3, pp. 442-451, Feb. 1995.

[37] H. Abdi, " Principal component analysis.,", edition 2, pp. 433-459, 2010.

[38] Kevin P.Murphy, "Machine Learning: A Probabilistic Perspective," Cambridge, Mass.: MIT Press, pp. 387-395, 2012.

[39] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophys. J.*, vol. 80, no. 1, pp. 505-515, Jan. 2001.

[40] F. Tama and Y. H. Sanejouand, "Conformational change of proteins arising from normal mode calculations," *Protein Eng*, vol. 14, no. 1, pp. 1-6, Jan. 2001.

[41] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, and A. R. Ortiz, "An analysis of core deformations in protein superfamilies," *Biophys. J.*, vol. 88, no. 2, pp. 1291-1299, Feb. 2005.

CHAPTER 4. STRUCTURAL MODELING OF FRUCTOSE BISPHOSPHATE ALDOLASE AND TRIOSE PHOSPHATE ISOMERASE INTERACTION – A MECHANISTIC PERSPECTIVE

Manuscript prepared for submission to a peer-reviewed scientific journal

Ataur R. Katebi and Robert L. Jernigan

Abstract

Fructose bisphosphate aldolase (FBA) and triosephosphate isomerase (TIM) are the fourth and the fifth enzymes in the glycolysis pathway. FBA cleaves the six-carbon fructose 1, 6-bisphosphate (FBP) into two three-carbon components – dihydroxy acetone phosphate (DHAP) and glyceraldehyde 3-phosphate (GAP). GAP is the correct substrate for the subsequent enzyme GAPDH, but DHAP is not so. DHAP is shunted to TIM as its substrate where it is converted to a second molecule of GAP, which is the substrate for GAPDH. FBA and TIM are both alpha/beta barrel proteins that are highly structurally similar, having an RMSD of 4.8Å for their cores. Their functional loops are also aligned after the superposition of their cores. Moreover, inspection of the sequences of these two proteins across different species shows that the C-terminus of the functional loop 5 in the FBA structure carries a 'phosphate gripper' motif and the tip of the functional loop 6 in the TIM structure has a similar motif. These motif-carrying loops are highly mobile, and each adopts alternative open and closed conformations, before and after substrate binding. When open, the functional loops are suitable for substrate recruitment – FBP for FBA and DHAP for TIM. On the other hand, analyses of the dynamics of each of the FBA and TIM proteins show that the functional loops (front loops 5, 6, and 7 in FBA, and 6, 7, and 8 in TIM) within each structure move in a highly coordinated ways. These are clear indications that the

dynamics of the structural components that form the catalytic microenvironment are similarly synchronized in the two enzymes. Considering the architectural similarity and distinctness, the functional loop coordinations within and between FBA and TIM structures, the presence and the placement of the 'phosphate gripper' on one of the functional loops in each structure, this provides significant indication that an FBA-TIM pair could function as a coupled and coordinated machine.

4.1 Introduction

In this chapter we will discuss different structural and mechanical issues that relate to interactions between FBA and TIM proteins. FBA and TIM are the fourth and the fifth enzymes in the glycolysis pathway which is present in all prokaryotes and eukaryotes. Along the glycolysis pathway, a six carbon sugar molecule goes through ten steps. At each step, a specific chemical reaction is catalyzed by a specific enzyme [7]. It has long been thought that sequential enzymes along metabolic pathway are likely to be arranged adjacently within the cell. Figure 1 shows a partial glycolysis pathway for the interactions retrieved from the KEGG database [8;9] and the Biogrid database [10]. In this partial pathway, there are more interactions among participant enzymes than those shown in Fig. 1 of Chapter 1. This indicates that the glycolysis pathway is more complex than the simple linear process described in Chapter 1. In this process, FBA picks up its substrate FBP from PFK, cleaves it into two products GAP and DHAP. Subsequently, TIM converts DHAP into GAP. Both FBA and TIM release their products GAP molecule to GAPDH as its substrate.

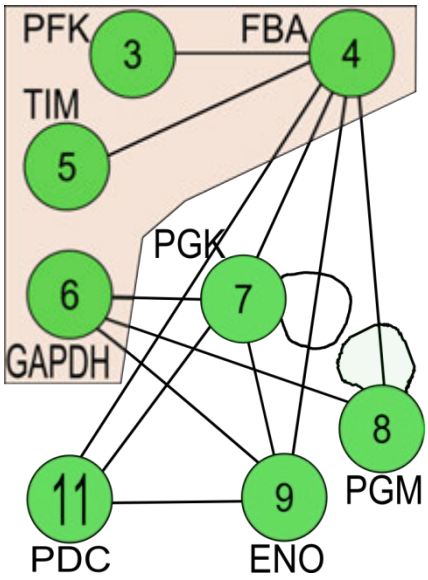


Figure 25. Interactions reported between FBA and TIM in the partial glycolysis pathway of eight proteins involved in steps 3, 4, 5, 6, 7, 8, 9 and pyruvate decarboxylase (PDC) that connects the glycolysis cycle with the tricarboxylic acid (TCA) cycle. The loops indicate self-interactions. The enzymes in the shaded region are the four consecutive enzymes that will be further used to model interaction mechanism in Chapter 6. It is noteworthy that FBA is linked with TIM and that neither FBA nor TIM is connected directly with GAPDB, suggesting some asymmetry in the mechanism. Acronyms are the same as those given in Fig. 1 of Chapter 1.

The interaction between FBA and TIM was identified by Affinity Capture Mass Spectrometry method and computational methods have further shown this to be a high-confidence interaction [31]. Here, we compare the sequences, structures, functions, and dynamics of these two enzymes. We have detailed the dynamics of how FBA and TIM change upon oligomerization in Chapters 2 and 3, respectively. In this chapter, we explore the feasibility of the interaction between FBA and TIM.

4.1.1 Comparing FBA and TIM Structures

Each subunit of FBA and TIM has a central beta/barrel consisting of eight β-strands, β1 ~ β8, where the barrel is surrounded by a helical ring that consist of eight α-helices, α1 ~ α8. In case of FBA, helix 8 is divided into two separate shorter helices – helix α8A and helix α8B, linked by a coil region. Also, FBA has an extra helix, α0 that covers the bottom of the barrel. There are eight

front loops and eight back loops – front loops running from strand to helix and back loops running from helix to strand. Figures 2A and 2B show the architectures of subunits of FBA and TIM with their structural and functional components identified. Table I shows the lengths and positions of the secondary structures in FBA and TIM.

Active forms of FBA and TIM are oligomers. Figures 3A and 3C show the organization of FBA and TIM dimeric structures from *S.cerevisiae*, respectively. Each dimer is formed from two subunits by the interactions at the type 1 interface. Figures 3B and 3D show the tetrameric organization of FBA and TIM from *T.aquaticus* and *P.woesei*, respectively. Each tetramer is formed by two additional type 2 interfaces. Type 1 and type 2 interfaces in each case have been defined and described in details for FBA and TIM structures in Chapters 2 and 3, respectively. The hinge region along the interface controls the global motions of each structure. The dimerization region (type 1 interface) in each structure consists of complementary structural components from each subunit. The global motions of the protein and the local motions of the functional loops provide the required open/closed conformations: in the open conformation the substrate can enter the catalytic cavity. In the closed conformation, the catalytic activity takes place. In FBA, the three functional loops marked as front loops 5, 6, and 7 in Table I make an inward excursion from open to closed conformation and cover the catalytic pocket. In TIM, the corresponding functional loops marked as front loops 6, 7, and 8 in the table make the similar inward excursion to cover the catalytic pocket.

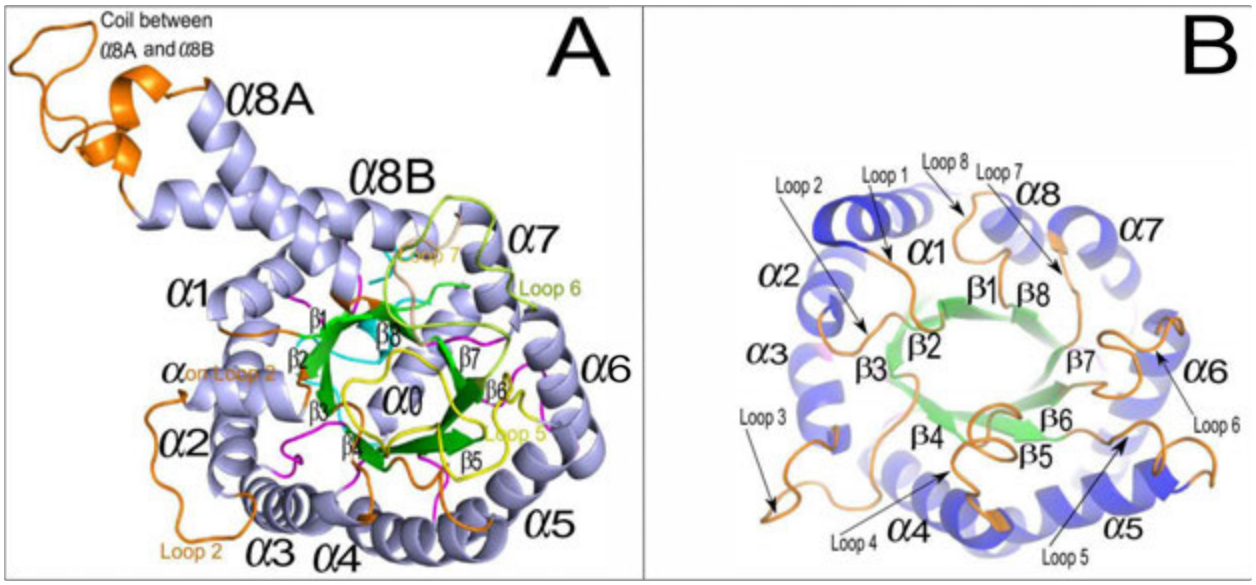


Figure 26. Comparison of the FBA and TIM subunit architectures. (A) Different structural components of an FBA monomer based on the *E.coli* FBA structure (PDB 1B57 – a dimer). Eight β-strands, labeled β1 ~ β8 and colored green, form the central barrel and eight helices, labeled α1 ~ α8 and colored blue, surround this β-barrel. Loops connecting the helices to strands located on the N-terminal side are colored magenta. Loops connecting the strands to helices on the C-terminal side are colored orange or yellow (Loop 5) or light green (Loop 6) or gray (Loop 7). (B) Different structural components of a TIM monomer from *S.cerevisiae* (PDB 1YPI – a dimer). Similar to the FBA structure in panel A, eight strands, labeled β1 ~ β8 and colored green, form the central barrel and eight helices α1 ~ α8 surrounds this barrel. Back loops are not shown. Front loops connecting the strands with the helices are labeled and colored orange. As can be noticed, FBA helices and front loops are comparatively longer than those in TIM.

Although FBA and TIM maintain similar (α/β)-barrel architectures their functions are nonetheless quite different from each other. They achieve this functional difference by the differences in their catalytic microenvironments. The details of the construction of the two catalytic sites can be found in section 2.1.5 of Chapter 2 for FBA and section 3.1.3 of Chapter 3 for TIM. Each subunit of FBA has an active site cavity where catalysis – the cleaving of its substrate FBP into DHAP and GAP, can take place. The opening/closing of this cavity and the catalytic activity of this enzyme is controlled by three functional loops (loop 5, loop 6, and loop 7). Each subunit of TIM likewise has a catalytic cavity where isomerization between DHAP and

GAP takes place. Very similar to the FBA structure, the cavity opening/closing and catalysis are controlled by three functional loops (loops 6, 7, and 8).

Table I. Positions of Secondary Structure Elements in the Sequences of *E.coli* FBA and *S.cerevisiae* TIM

Type of Secondary Structure	Residue Indices in the Protein Sequence			
	E.coli FBA		*S.cerevisiae* TIM	
	Indices	Length	Indices	Length
N-terminus	1 – 15	15	1 – 4	4
N-terminal Helix 0	16 – 26	11		
N-terminal Loop 0	27 – 30	4		
Helices				
Helix 1	40 – 52	14	17 – 29	13
Helix 2	80 – 100	20	46 – 53	8
Helix 3	116 – 133	18	80 – 85	6
Helix 4	151 – 165	15	106 – 118	13
Helix 5	199 – 209	11	139 – 153	15
Helix 6	239 – 252	14	177 – 196	10
Helix 7	272 – 278	7	218 – 221	4
Helix 8A	291 – 305	15	240 – 245	6
Coli between helices 8A & 8B	306 – 330	25		
Helix 8B	331 – 352	22		
Back Loops				
Loop 1	53 – 55	3	30 – 35	6
Loop 2	100 – 102	4	54 – 58	5
Loop 3	134 – 139	6	86 – 89	4
Loop 4	166 – 169	4	119 – 121	3
Loop 5	210 – 215	6	154 – 159	4
Loop 6	253 – 259	7	197 – 205	9
Loop 7	279 – 283	5	222 – 227	6
Strands				
Strand 1	31 – 35	5	5 – 10	6
Strand 2	56 – 60	5	36 – 41	6
Strand 3	103 – 108	6	59 – 63	5
Strand 4	140 – 143	4	90 – 93	4
Strand 5	170 – 175	6	122 – 127	6
Strand 6	216 – 220	5	160 – 164	5
Strand 7	260 – 263	4	206 – 209	4
Strand 8	284 – 287	4	228 – 231	4
Front Loops				
Loop 1	35 – 39	3	11 – 16	6
Loop 2	61 – 79	20	42 – 45	4
Loop 3	109 – 115	7	64 – 79	6
Loop 4	144 – 150	7	94 – 105	12
Loop 5	176 – 198	23	128 – 138	11
Loop 6	221 – 238	18	165 – 176	12
Loop 7	264 – 271	8	210 – 217	18
Loop 8	288 – 290	3	228 – 239	12
C-terminus	353 – 358	6	246 – 248	3

Note: Significant differences in lengths of: helices 2, 3, and 6, coil between helices 8A and 8B, front loops 2, 4, 5, 6, 7 and 8

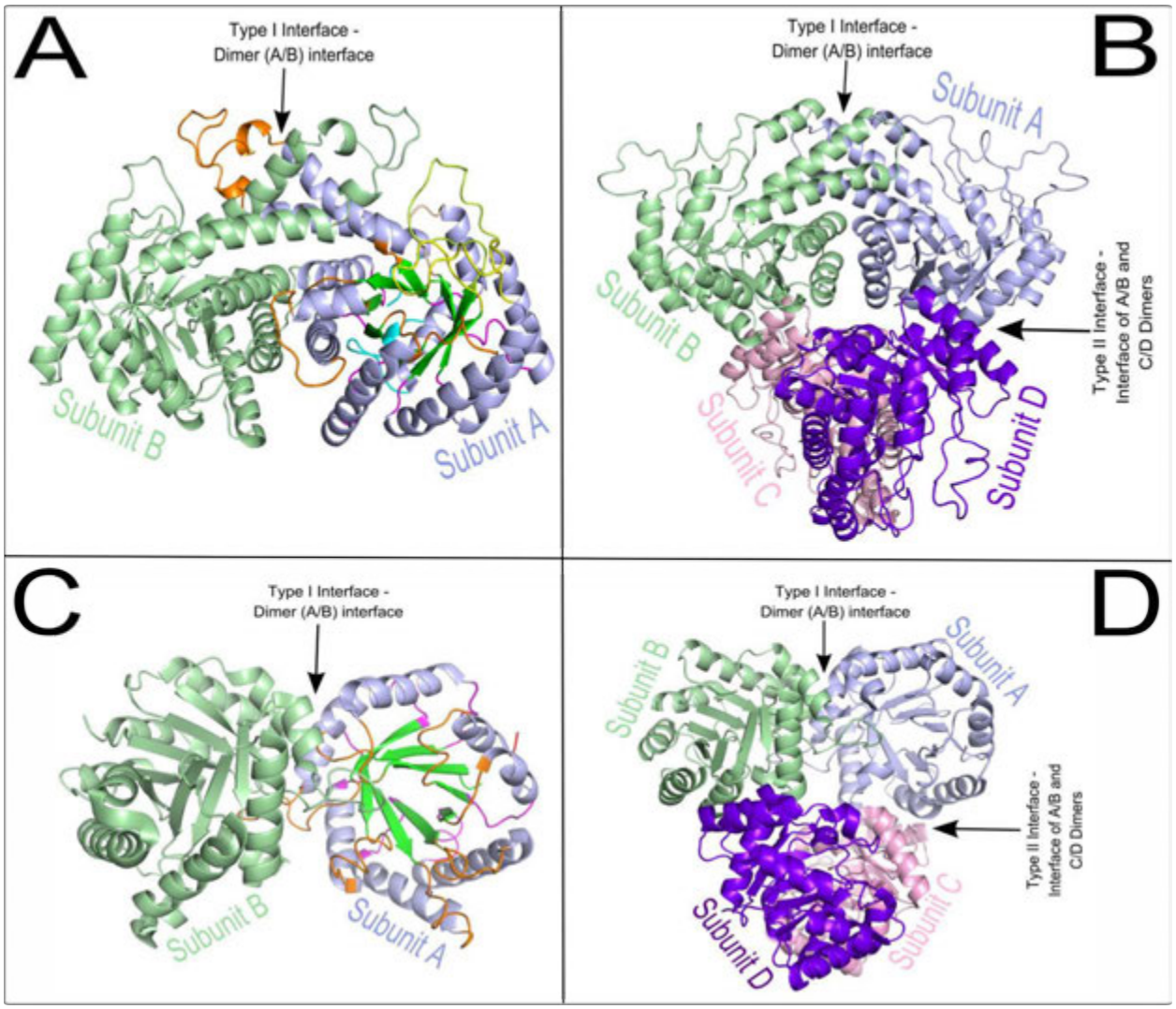


Figure 27. Oligomeric architectures of FBA and TIM. (A) An FBA dimer – two subunits bound together by the interactions along the type 1 interface region. (B) An FBA tetramer – two type 1 dimers join together to form the tetramer through the interactions along two type 2 interfaces. (C) TIM dimer – similar to the FBA dimer in panel A, two TIM subunits are bound together at a type 1 interface. (D) TIM tetramer – similar to an FBA tetramer, two type 1 TIM dimers are bound together by two type 2 interfaces.

Fold Similarity and Functional Loop Correspondence between FBA and TIM

FBA and TIM have different sizes: each subunit of *S.cerevisiae* FBA has 358 residues whereas each subunit of TIM from the same organism has only 247 residues. The extra residues in FBA are mainly used to create some extra parts used in its dimerization. The differences that

are created in the construction of the dimerization interfaces cause difference in the structural mobility and thus facilitate the distinctive motions of the catalytic regions.

Moreover, the superposition of the cores – the central β-barrel surrounded by the eight helices in each structure, of these two enzyme structures yields an RMSD value of 4.88Å which is quite small compared to their very low sequence identity (7.28%) and similarity (11.42%) (see Fig. 4). It also aligns the functional loops of FBA with the corresponding functional loops of TIM. An investigation of the sequences reveals that the tip of loop 6 in TIM structure includes the 'phosphate gripper' motif [25].TIM uses this motif to trap its substrate from the vicinity. The N-terminus of loop 5 of FBA also has such a motif that the enzyme could use to channel its substrate from the vicinity to its catalytic pocket.

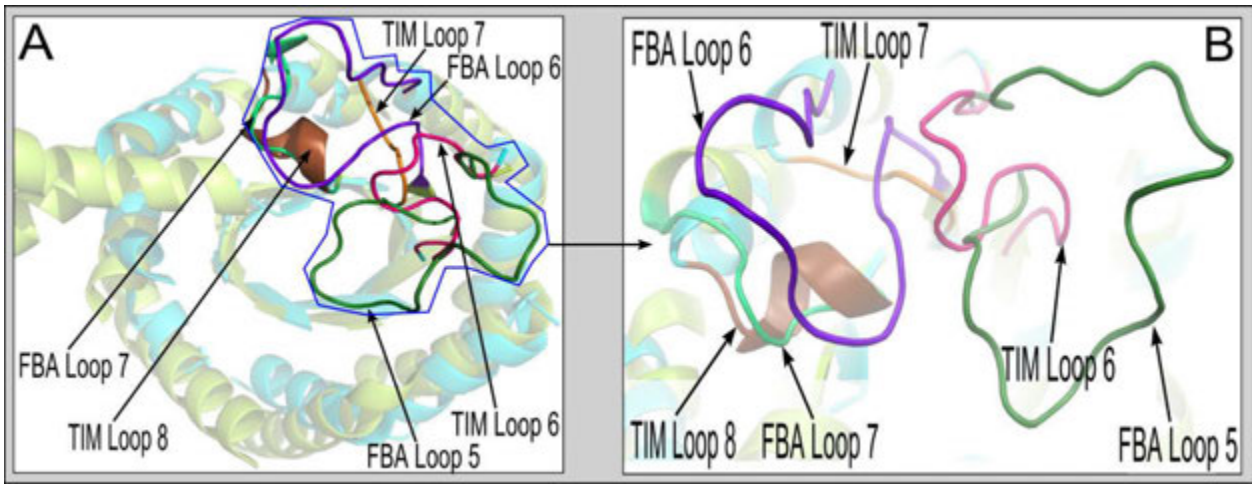


Figure 28. Comparison of the FBA and TIM cores. (A) FBA and TIM core alignment gives an RMSD between them of 4.88 Å. Strand 1 of TIM aligns with the strand 8 of FBA. Thus FBA functional loop 5, 6, and 7 align with TIM functional loops 6, 7, and 8, respectively. Colors for the cores: yellow – FBA, cyan – TIM. Colors for the functional loops: green – FBA loop 5, magenta – TIM loop 6, olive – FBA loop 6, yellow – TIM loop 7, light green – FBA loop 7, dark gray – TIM loop 8. (B) Enlarged view of the aligned functional loops from panel A.

4.1.2 Modeling the Mechanism of FBA-TIM Interaction

It has been shown in many experiments that the function of a protein depends not only on its sequence but also on its structure and the associated dynamics. The residues that are important for maintaining the structure of the protein could be different from the residues that contribute to its functionality. Oligomerization of a protein can be important for both its dynamics and functionality. A protein can be in either monomeric form or in higher oligomeric form that could be either a homocomplex or a heterocomplex. Homocomplexes tend to be permanent and optimized whereas the heterocomplexes can be either permanent or nonobligatory. Nonobligatory complexes require that the component proteins must be capable of existing independently from each other , i.e., they may be folded and not depending on the environmental conditions [11]. Interfaces of homocomplexes such as dimeric protein structures have hydrophobic parts different from interior hydrophobicity [5]. Chothia, C. and Janin, J. emphasize hydrophobicity and complementarity in such protein-protein recognition. Complementarity of the binding interface residues helps two proteins to associate by forming hydrogen bonds and van der Waal's contacts between residues. Better complementarity helps these interface residues pack more closely together [22]. Once the two proteins bind together in their complementary regions, the stability of the newly formed complex also depends on the hydrophobicity of the binding interfaces for its stability [12]. The hydrophobicity of the interface region is correlated with the buried surface area. While the complementarity plays a selective role in deciding which proteins may associate, the hydrophobicity is the major factor in stabilizing protein-protein association [15]. Proteins may go through some conformational changes upon binding and these conformational changes are related to the size of the recognition site [16]. Onuchic *et al* postulated from their simulations that binding processes have funneled landscapes similar to

folding processes. They also concluded that binding mechanism is robust and is governed by protein topology as it does in case of folding. Their model showed that the degree of topological frustration determines whether binding will occur between two unfolded or folded chains. These results emphasize the previous findings that protein dynamics and plasticity are essential for bimolecular recognition and the binding scenarios are much more diverse than previously imagined [14].

Although the physical principles of protein-protein recognition are fairly well understood, their application for predicting the complexes between two proteins is not as well established [6]. Modeling the interaction between two proteins from the structural point of view requires quantitative definitions of the binding energy and complementarity derived from structures. Protein-protein binding energy is often taken to be the sum of the energetic and entropic changes contributions of the individual amino acid residues [13]. Some computational protein-protein docking methods such as ClusPro [26], Z-Dock [27], Rosetta Dock [28], HADDOCK [29], FireDock [30], etc. have been developed based on these concepts. These methods try to find good models for a complex between the interacting proteins based on the structural information. Some attempt to incorporate some aspects of protein dynamics such as side chain conformational flexibility and large scale protein conformational switches. The integration of the full scale protein functional motions into these models has been a challenge with little success, because of the enhanced complexity of the problem. Recent results from our laboratory have shown that even crude estimates of whole protein entropies can improve the selection of docked forms significantly.

Partly this succeeds because the dominant motions in protein dynamics are the large domain motions that require proper accounting of the simultaneous motions of the whole protein. Their

motions play crucial roles in how they recruit their substrate, perform their function, and interact with other proteins. In Chapter 2, we investigated how the dynamics of FBA changes across the interface and functional regions with its oligomerization. In Chapter 3, we performed similar investigation on TIM. Oligomerization helps enzymes not only to attain stability but also to achieve the appropriate dynamics essential to carrying out their functions.

This chapter investigates the mechanistic feasibility of the interaction of these two enzymes whose oligomerization and related dynamics we studied in the last two chapters. Here we employ Elastic Network Models (ENMs) [1] to investigate and compare the dynamics of FBA and TIM structures from *S.cerevisiae*. We compare the fluctuations of the functional loops within FBA and TIM. We have shown in Figs. 4A and 4B that the functional loops are aligned after the superimposition of the subunit structures of these two enzymes – FBA loops 5, 6, and 7 align with TIM loops 6, 7, and 8. We also investigate how the motions of the functional loops in one enzyme correlate with those of the functional loops in the other enzyme. These provide a way to investigate the synchrony of the dynamics of these two proteins.

4.2 Results

Correlations of Motions within FBA and TIM Functional Loops

Figure 5 compares the directions of motion between the coorresponding functional loop pairs in the FBA and TIM structures. FBA maintains the highest correlation between loops 6 and 7 while TIM does this for loops 7 and 8. We also notice that the coordination between loops 5 and 6 in FBA is higher than the coordination between loops 6 and 7 in TIM. But in the case of loop pair 5 and 7 in FBA and loop pair 6 and 8 in TIM, TIM maintains a better correlation. The high correlation values between each pair of functional loops within each of the FBA and TIM

structures clearly indicate that the catalytic process in each catalytic pocket is performed by the concerted motions of the functional loops.

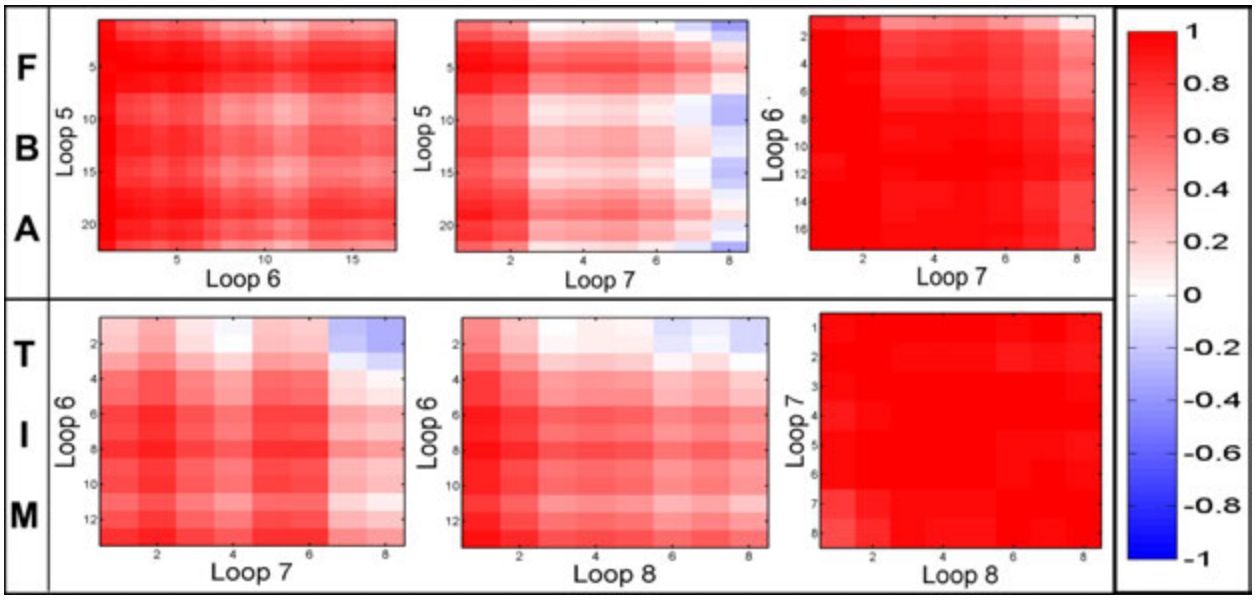


Figure 29. Correlations of motions between the corresponding loops in FBA and TIM. (Top) Correlations of motions between the functional loop pairs within FBA – there is high correlation between loop 5 and loop 6; similarly, high correlation exists between loop 5 and loop 7, and also between loop 6 and loop 7. Closed FBA structure with PDB Id 1B57 is used. (Bottom) Correlations of motions between the functional loop pairs in the TIM structure – loop 6 vs loop 7, loop 6 vs loop 8, and loop 7 vs loop 8. Closed TIM structure with PDB Id 7TIM is used.

Correlations of Motions between FBA and TIM Functional Loops

Figure 6 shows the overlap of modes of motion and the cumulative overlap of modes of motions for the three FBA/TIM functional loop pairs – 5/6, 6/7, and 7/8. Figs. 6A, 6B, and 6C show that the overlap of motions in mode 1 are 0.78, 0.86, and 0.95, between the three corresponding functional loop pairs – FBA loop 5/TIM loop 6, FBA loop 6/TIM loop 7, and FBA loop 7/TIM loop 8, respectively. This indicates that the global motions of the structures bring the functional loops between the structures in high synchrony. For FBA loop 5/TIM loop 6 pair, the overlap in mode 2 is 0.83 and this overlap between mode 3 (of FBA) and mode 5 (of TIM) is 0.70. For FBA loop 7/TIM loop 8 pair, the overlap in mode 4 is 0.82, and this overlap in

mode 3 and mode 2 is 0.82. No other values are as significant. This indicates that FBA loop 5/TIM loop 6 and FBA loop 7/TIM loop 8 also maintain high coordination in several modes of motions. We know that the catalytic residues in FBA structure are prominently located on functional loops 5 and 7 as shown in Table V of Chapter 2. Also the functional loops 6 and 8 in TIM coordinate the rate of substrate binding/release i.e. turnout rate because of the existence of the 'phosphate gripper' on loop 6 and the phosphate binding specificity residue on loop 8. Therefore we can conclude that the catalytic activity in the FBA structure is highly coordinated with the turnout rate of the TIM structure.

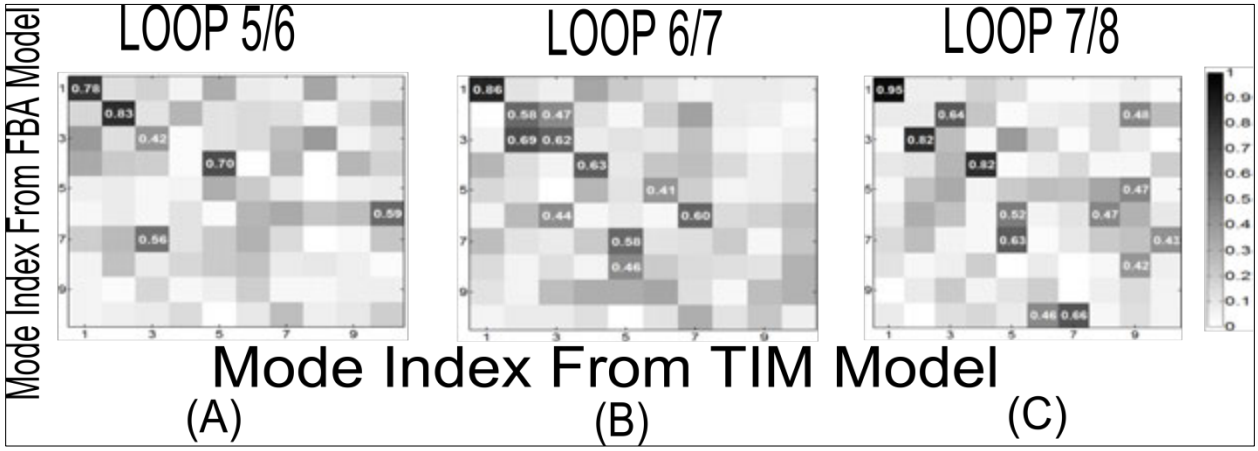


Figure 30. Overlap in directions of modes of the functional loops between FBA and TIM, both are in the closed conformations (PDB Id for FBA structure is 1B57 and PDB Id for TIM structure is 7TIM). Overlap of the modes of motions between FBA/TIM functional loop pairs – as they aligned shown in Figs. 4A and 4B – (A) FBA loop 5/TIM loop 6, (B) FBA loop 6/TIM loop 7, (C) FBA loop 7/TIM loop 8.

Another important issue with these loops is the presence of the 'phosphate gripper' on one of the functional loops – the C-terminus of FBA loop 6 and the tip of TIM loop 6. High synchronicity between loops 5 and 6 within the FBA structure (top panel of Fig. 5) and high coordination between the FBA loop 5 and the TIM loop 6 further indicate that substrate recruitment in each of these structures may have some coordination.

Interestingly, for each pair of the functional loops, the more local modes of motions between FBA and TIM structures have low overlaps as shown in Figs. 6A, 6B, and 6C. This indicates that only the most global motions play the critical role in keeping the dynamics of the two structures in harmony.

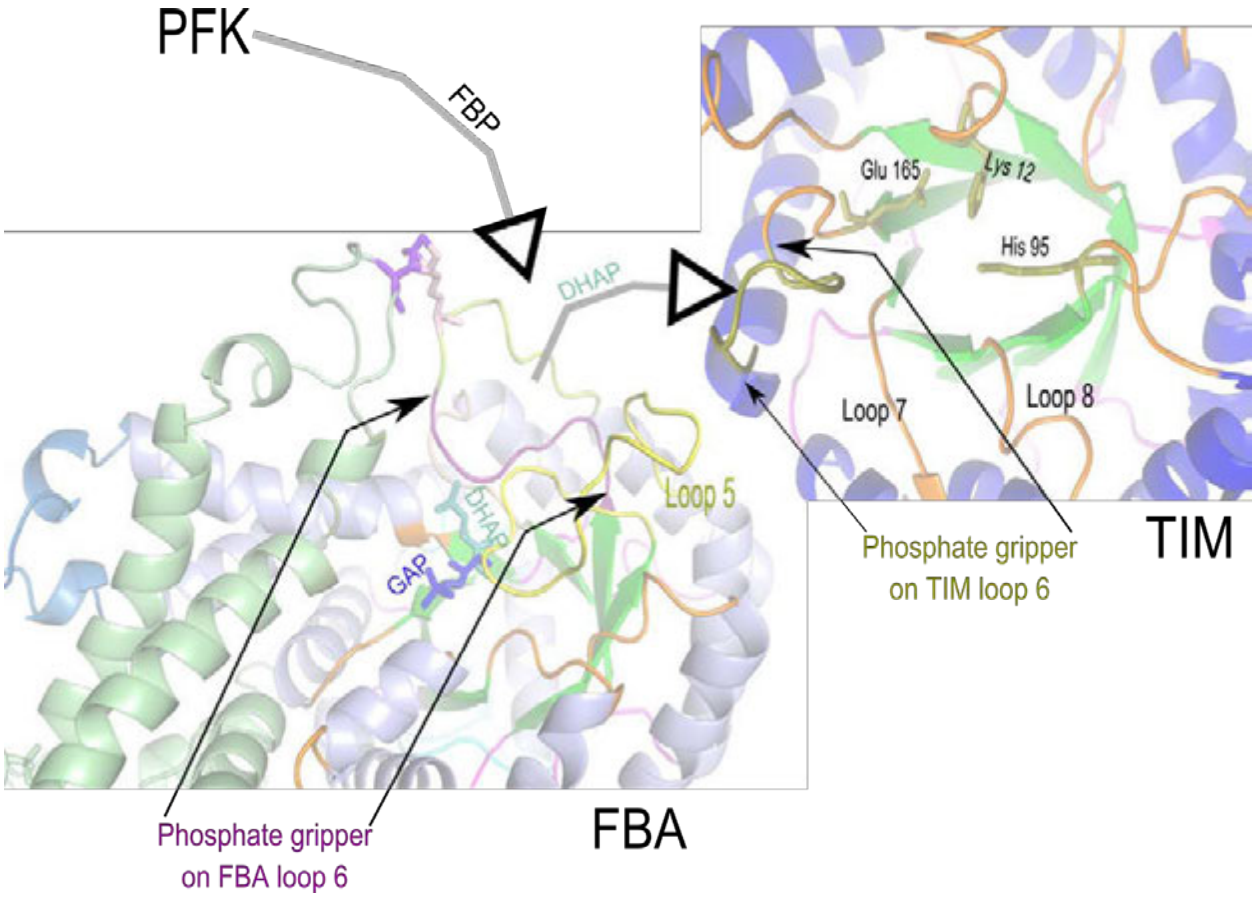


Figure 31. A putative model for substrate transfer from FBA to TIM. The acronyms are the same as defined in Fig. 1 of Chapter 1. Each enzyme structure may use its 'phosphate gripper' to recruit its substrate from the product of the previous enzyme structure – FBA can get it from PFK whereas TIM can get it from FBA. In the open conformation, FBA and TIM might use their 'phosphate grippers' to hunt for their substrate from the vicinity. In the closed conformation, FBA cleaves its FBP into two components – GAP and DHAP. While transitioning from closed to open, FBA removes the products from its active site so that DHAP is within the reach of the 'phosphate gripper' of open loop 6 of the neighboring TIM.

4.3 Discussion and Conclusion

Both FBA and TIM maintain a high correlation between the functional loop pairs within each structure. The patterns of correlations between functional loop pairs within each structure are also similar. Moreover, the overlaps between the aligned functional loop pairs from the FBA and TIM structures are quite high. This indicates that the coordination of the dynamics of the structural components forming the catalytic pocket within FBA is matched with the coordination of the similar components within TIM. This is a hallmark of some type of synchronization in the functional mechanism between FBA and TIM. Figure 7 shows a model for such a coupling between FBA and TIM. After the cleaving of FBP in the catalytic pocket of FBA, the cleaved component DHAP still attached with the 'phosphate gripper' of FBA is ejected as the FBA loop 6 swings to its open conformation. DHAP reaches the proximity of the 'phosphate gripper' on TIM loop 6 and is pulled into the TIM catalytic pocket as loop 6 swings to its closed conformation. The proposed model in Fig. 8 can explain substrate transfer between FBA and TIM in a synchronized fashion. High correlations among the functional loops within FBA and TIM, and the high overlaps of functional loop motions between FBA and TIM, support such a mechanism.

4.4 Methods and Materials

Modeling *S.cerevisiae* FBA Enzyme

Since the structure for *S.cerevisiae* FBA is not available in the PDB database, we modeled this structure. The details of the modeling can be found in section 2.4.2 of Chapter 2.

4.4.2 Modeling Dynamics

We have applied coarse-grained Elastic Network Model (ENM) to compute the dynamics of the proteins. The details of the ENM can be found in section 2.4.3 of Chapter 2. We computed overlaps of the motions of the different components of the structure. The models to compute these overlaps can be found in 3.4.3 of Chapter 3.

Authors' contributions

ARK and RLJ both contributed to the design, execution and writing of this work.

Bibliography

[1] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model," *Biophys. J.*, vol. 80, no. 1, pp. 505-515, Jan. 2001.

[2] A. Skliros, M. T. Zimmermann, D. Chakraborty, S. Saraswathi, A. R. Katebi, S. P. Leelananda, A. Kloczkowski, and R. L. Jernigan, "The importance of slow motions for protein functional loops," *Phys. Biol.*, vol. 9, no. 1, p. 014001, Feb. 2012.

[3] O. Kurkcuoglu, R. L. Jernigan, and P. Doruker, "Loop motions of triosephosphate isomerase observed with elastic networks," *Biochemistry*, vol. 45, no. 4, pp. 1173-1182, Jan. 2006.

[4] O. Kurkcuoglu, Z. Kurkcuoglu, P. Doruker, and R. L. Jernigan, "Collective dynamics of the ribosomal tunnel revealed by elastic network modeling," *Proteins*, vol. 75, no. 4, pp. 837-845, Jun. 2009.

[5] S. Jones and J. M. Thornton, "Protein-protein interactions: a review of protein dimer structures," *Prog. Biophys. Mol. Biol.*, vol. 63, no. 1, pp. 31-65, 1995.

[6] P. Zielenkiewicz and A. Rabczenko, "Methods of molecular modelling of protein-protein interactions," *Biophys. Chem*, vol. 29, no. 3, pp. 219-224, Apr. 1988.

[7] L. A. Fothergill-Gilmore and P. A. Michels, "Evolution of glycolysis," *Prog. Biophys. Mol. Biol.*, vol. 59, no. 2, pp. 105-235, 1993.

[8] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Res.*, vol. 40, no. Database issue, p. D109-D114, Jan. 2012.

[9] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27-30, Jan. 2000.

[10] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. Database issue, p. D535-D539, Jan. 2006.

[11] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proc. Natl. Acad. Sci U. S. A*, vol. 93, no. 1, pp. 13-20, Jan. 1996.

[12] C. Chothia and J. Janin, "Principles of protein-protein recognition," *Nature*, vol. 256, no. 5520, pp. 705-708, Aug. 1975.

[13] A. Horovitz, "Non-additivity in protein-protein interactions," *J Mol. Biol.*, vol. 196, no. 3, pp. 733-735, Aug. 1987.

[14] Y. Levy, P. G. Wolynes, and J. N. Onuchic, "Protein topology determines binding mechanism," *Proc. Natl. Acad. Sci U. S. A*, vol. 101, no. 2, pp. 511-516, Jan. 2004.

[15] J. Janin and C. Chothia, "The structure of protein-protein recognition sites," *J. Biol. Chem.*, vol. 265, no. 27, pp. 16027-16030, Sept. 1990.

[16] C. L. Lo, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *J. Mol. Biol.*, vol. 285, no. 5, pp. 2177-2198, Feb. 1999.

[17] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223-230, Jul. 1973.

[18] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics.*, vol. 20, no. 1, pp. 45-50, Jan. 2004.

[19] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics.*, vol. 20, no. 1, pp. 45-50, Jan. 2004.

[20] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda, "PIPER: an FFT-based protein docking program with pairwise potentials," *Proteins*, vol. 65, no. 2, pp. 392-406, Nov. 2006.

[21] D. Kozakov, D. R. Hall, D. Beglov, R. Brenke, S. R. Comeau, Y. Shen, K. Li, J. Zheng, P. Vakili, I. C. Paschalidis, and S. Vajda, "Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19," *Proteins*, vol. 78, no. 15, pp. 3124-3130, Nov. 2010.

[22] L. Pauling, "Molecular basis of biological specificity," *Nature*, vol. 248, pp. 769-771, Apr. 1974.

[23] F. Tama and Y. H. Sanejouand, "Conformational change of proteins arising from normal mode calculations," *Protein Eng*, vol. 14, no. 1, pp. 1-6, Jan. 2001.

[24] A. Leo-Macias, P. Lopez-Romero, D. Lupyan, D. Zerbino, and A. R. Ortiz, "An analysis of core deformations in protein superfamilies," *Biophys. J.*, vol. 88, no. 2, pp. 1291-1299, Feb. 2005.

[25] J. R. Knowles, "Enzyme catalysis: not different, just better," *Nature*, vol. 350, no. 6314, pp. 121-124, Mar. 1991.

[26] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics.*, vol. 20, no. 1, pp. 45-50, Jan. 2004.

[27] R. Chen and Z. Weng, "Docking unbound proteins using shape complementarity, desolvation, and electrostatics," *Proteins*, vol. 47, no. 3, pp. 281-294, May 2002.

[28] S. Lyskov and J. J. Gray, "The RosettaDock server for local protein-protein docking," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, p. W233-W238, Jul. 2008.

[29] C. Dominguez, R. Boelens, and A. M. Bonvin, "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information," *J Am. Chem Soc.*, vol. 125, no. 7, pp. 1731-1737, Feb. 2003.

[30] N. Andrusier, R. Nussinov, and H. J. Wolfson, "FireDock: fast interaction refinement in molecular docking," *Proteins*, vol. 69, no. 1, pp. 139-159, Oct. 2007.

[31] S.R. Collins, P. Kemmeren, *et al*, "Toward a Comprehensive Atlas of the Physical Interactome of *Saccharomyces cerevisiae*," *Molecular and Cellular Proteomics*, vol. 6, no. 3, pp. 439-450, Mar. 2007.

[32] M.T. Zimmermann, S.P. Leelananda, A. Kloczkowski, and R.L. Jernigan, "Combining statistical potentials with dynamics-based entropies improves selection from protein decoys and docking poses," Journal of Physical Chemistry B, vol. 116, no. 23, pp. 6725-6731, Jun. 2012.

CHAPTER 5. STRUCTURAL INTERPRETATION OF PROTEIN-PROTEIN INTERACTION NETWORK

This is a published manuscript in the peer reviewed scientific journal, *BMC Structural Biology*

Ataur R. Katebi, Andrzej Kloczkowski, Robert L. Jernigan

Abstract

Background

Currently a huge amount of protein-protein interaction data is available from high throughput experimental methods. In a large network of protein-protein interactions, groups of proteins can be identified as functional clusters having related functions where a single protein can occur in multiple clusters. However experimental methods are error-prone and thus the interactions in a functional cluster may include false positives or there may be unreported interactions. Therefore correctly identifying a functional cluster of proteins requires the knowledge of whether any two proteins in a cluster interact, whether an interaction can exclude other interactions, or how strong the affinity between two interacting proteins is.

Methods

In the present work the yeast protein-protein interaction network is clustered using a spectral clustering method proposed by us in 2006 and the individual clusters are investigated for functional relationships among the member proteins. 3D structural models of the proteins in one cluster have been built – the protein structures are retrieved from the Protein Data Bank or predicted using a comparative modeling approach. A rigid body protein docking method (Cluspro) is used to predict the protein-protein interaction complexes. Binding sites of the

docked complexes are characterized by their buried surface areas in the docked complexes, as a measure of the strength of an interaction.

Results

The clustering method yields functionally coherent clusters. Some of the interactions in a cluster exclude other interactions because of shared binding sites. New interactions among the interacting proteins are uncovered, and thus higher order protein complexes in the cluster are proposed. Also the relative stability of each of the protein complexes in the cluster is reported.

Conclusions

Although the methods used are computationally expensive and require human intervention and judgment, they can identify the interactions that could occur together or ones that are mutually exclusive. In addition indirect interactions through another intermediate protein can be identified. These theoretical predictions might be useful for crystallographers to select targets for the X-ray crystallographic determination of protein complexes.

5.1 Background

Because of the use of high throughput experimental methods such as yeast two-hybrid screening[1], the number of reported protein-protein interactions (PPI) has increased dramatically. To extract meaningful information from this interaction data set, clustering of the interacting proteins is an established method. Patra *et al.* [2] have shown that functionally significant clusters can be extracted from the dominant eigenvalues of a modified contact matrix known as the Kirchhoff matrix. Sen *et al*.[3] used an eigenmode analysis (a type of spectral clustering) to cluster the interacting proteins.

The BioGrid database has published different versions of yeast protein interaction data with increasing numbers of proteins and interactions[4]. Some limited attempts have been made to construct spatial interaction clusters from this data. With early results showing that such clusters have functional relationships, such results may help to predict undiscovered interactions among proteins in the same cluster[3]. However the protein interaction data obtained from high-throughput screening methods such as the yeast two-hybrid method[1] and affinity purification techniques[5] are highly error-prone. Approximately, 30–60% false positives and 40–80% false negatives have been estimated for these methods [6;7]. Therefore predicting new interactions or drawing any conclusions from this interaction dataset requires validation of the interactions. Another complementary source of information about the proteins is their individual structures. If there were sufficient known structures of the protein-protein pairs they could provide direct validation of the interactions. However, the number of such known structures remains small, and certainly nowhere near the number of interacting pairs that have been reported. But there are relatively large numbers of individual protein structures. Those, together with improvements in docking methods make it possible to begin investigating the likelihood of forming individual three dimensional pairs of structures[8]. Looking at the 3D structure of each protein, especially the binding sites, in an interacting cluster can reveal information that can aid in validating the pair-wise interactions. Some questions that we set out to investigate here are:

1. Whether two proteins prefer to interact?

2. If more than two proteins purportedly interact with the same protein, can they interact concurrently by binding two separate regions of the protein, or does one exclude the other because their binding sites substantially overlap?

3. What are the relative binding strengths of proteins within a cluster?

We choose the yeast protein-protein interaction network from the online database BIOGRID(www.thebiogrid.org)[4]. The number of distinct proteins and interactions in the dataset has increased manyfold since the analysis by Sen *et al*.[3]. The current dataset (version 2.0.55) has over five thousand proteins and more than 145,000 interactions.

5.2 Methods

We applied an eigenmode analysis to cluster the protein interaction network. We formed the Kirchhoff matrix[2] M; the interaction matrix M:

$$M_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ interact} \\ 0 & \text{if } i \text{ and } j \text{ do not interact} \\ -\sum_{k=1,k\neq j}^{n} M_{ik} & i = j \end{cases} \quad (1)$$

Then, we performed eigenmode analysis of this matrix M. This definition automatically leads to a singular matrix (i.e. the determinant of the matrix M is zero) that must be analyzed with Singular Value Decomposition[3].

Singular value decomposition (SVD)

We calculated all eigenvalues and eigenvectors of the connectivity matrix by applying the SVD subroutine available in the LAPACK library[9].

If A is any matrix of size m×n (with m>=n), then A can be written as a product of three matrices:

$$A = U\Lambda V^T \quad (2)$$

where Λ is the square matrix of size n×n containing nonnegative values $\lambda_1, \lambda_2, \ldots, \lambda_n$ along the diagonal and zeros off diagonal, and U and V are two matrices of sizes m×n and n×n, respectively, having orthogonal columns, i.e.

$$\sum_{i=1}^{m} U_{ik}U_{im} = \delta_{km} \quad \text{and} \quad \sum_{i=1}^{n} V_{ik}V_{in} = \delta_{kn} \tag{3}$$

The Kirchhoff matrix M can be written as

$$M = V\Lambda U^T \tag{4}$$

where Λ is the diagonal matrix containing eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of M and U is the matrix formed from eigenvectors of M. Thus, the elements M_{ij} of the contact matrix M can be expressed as

$$M_{ij} = \sum_{k=1}^{n} \lambda_k u_{ki} u_{kj} \tag{5}$$

where u_{ki} denotes the i^{th} component of the eigenvector corresponding to the k^{th} eigenvalue. Equation 4 is the eigenvalue expansion of the contact matrix. From Equation 5, it follows:

$$M_{ii} = \sum_{k=1}^{n} \lambda_k u_{ki}^2 \tag{6}$$

The eigenvalues with the smallest indices corresponding to the largest absolute values of λ make the largest contributions and smaller eigenvalues contribute successively less[3].

Cluster formation

For each eigenvalue there is a corresponding eigenvector. The significant components of an eigenvector comprise a cluster where each component corresponds to one protein. The components with an absolute value greater than 0.05 are assumed to be significant[3]. The clusters for larger eigenvalues are thus the interesting ones.

Interaction complex formation within a cluster

After the clusters are constructed, we need to choose a cluster to do structural analysis. Figure 1 shows three representative clusters (10, 14, and 15) for their moderate size. Out of these, we chose cluster 14 for further structural analysis. Then we attempt to predict the interaction complexes, predict new interactions, and predict whether multiple interactions could occur concurrently. The steps of this process are shown in the flowchart in Figure 2. In part (a) of the figure, an interacting partner protein structure is either retrieved from the Protein Data Bank (PDB)[10] (www.rcsb.org), or if there is no structure of the protein we predict the structure by comparative modeling. Figure 2(b) shows that once we have both structures of a putative interacting pair, we then use docking to predict the structure of the interaction complex.

Comparative modeling

To predict an interaction complex or predict a new interaction, we require the protein structures of both interacting proteins. If the structure of a protein is not available in the PDB, we use comparative modeling approaches [11;12]. To predict the structure of the protein, we have relied upon Zhang's I-TASSER server[12-14] (http://zhang.bioinformatics.ku.edu/I-TASSER), which gave the best protein models at the Critical Assessment of Structure Prediction (CASP 7 and CASP 8), a community-wide, worldwide experiment designed to obtain an objective assessment of the state-of-the-art in structure prediction[15-17]. The I-TASSER algorithm consists of three consecutive steps: threading, fragment assembly, and iteration. During threading, I-TASSER generates the template alignments by a simple sequence Profile-Profile Alignment approach constrained with the secondary structure matches. Fragment assembly is performed on the basis of threaded alignments and the target sequences are divided into aligned and unaligned regions. The fragments in the aligned regions are used directly from the template

structures and the unaligned regions are modeled with ab initio simulations. Clusters of decoys are generated with the use of a knowledge-based force field. The cluster centroids are generated by averaging the coordinates of all clustered decoys and ranked based on the structure density. In the iteration phase, the steric clashes of the cluster centroids are removed and the topology is refined. The conformations with the lowest energy are selected.

The I-TASSER server returns the best five models with a c-score attached for each model. Also it returns the top ten templates used in the threading. The c-score is a confidence score that I-TASSER uses to estimate the quality of the predicted model. The calculation of c-score is based on the significance of the threading template alignments and the convergence parameters of the structure assembly simulations. When selecting one of these models, we select the model that comes from the largest cluster and has the best c-score. C-score is in the range [-5,2], where a higher c-score value signifies a better model[14].

Docking

After we have both structures in an interacting pair we use docking to predict the protein complex formed in a protein-protein interaction. We use the Cluspro server[18-23] for docking the interacting proteins to predict the protein complex. Cluspro is the first fully automated web-based program for docking proteins and was one of the top performers at CAPRI (Critical Assessment of Predicted Interactions) rounds 1-12, the community-wide experiment devoted to protein docking[24]. The Cluspro server is based on a Fast Fourier Transform correlation approach, which makes it feasible to generate and evaluate billions of docked conformations by simple scoring functions. It is an implementation of a multistage protocol: rigid body docking, an energy based filtering, ranking the retained structures based on clustering properties, and finally, the refinement of a limited number of structures by energy minimization. The server

(http://cluspro.bu.edu/) returns the top models based on energy and cluster size. We select one of the returned models after considering the energy and the size of the cluster – preferring lower energies and larger cluster sizes. As the Cluspro server implements rigid body docking, when a partner protein in a complex is structurally flexible Cluspro is not so able to account for this flexibility.

5.3 Results

We perform the eigen-analysis on the yeast network version 2.0.40 (5,226 proteins and 114,754 interactions) and 2.0.41 (5,425 proteins and 121,664 interactions) and find that the number of zero eigenvalues are 6 and 3, respectively, which are very small compared to those from the yeast network Sen et al. previously used[3] (4,906 proteins, 19,037 interactions, and number of zero eigenvalues 46). This decrease in the number of zero eigenvalues is an indication of the completeness of the yeast network.

The proteins and their interactions in clusters 10, 14, and 15 are shown in Figure 1. We note that the number of neighbors for each protein in each of these clusters falls within a relatively small range. Those ranges are 278 – 288 for the proteins in cluster 10; 261 – 286 for the proteins in cluster 14; and 265 – 286 for the proteins in cluster 15.

We search the gene ontology database[25] for the functions of the proteins in each cluster and find that the proteins in each cluster have related functions usually. This is consistent with previous findings [3;26]. Table 1 shows the functions of each of the proteins in clusters 10, 14, and 15. The majority of the proteins in clusters 10 and 14 are cell cycle related; while cluster 15 is related to protein folding and protein degradation. We also attempt to determine the statistical confidence regarding the functional coherency of the clusters. We used

FunSpec(http://funspec.med.utoronto.ca/)[27], a web based cluster interpreter for yeast, to measure the functional coherency of the clusters (see Table 2). FunSpec assesses the degree of functional enrichment for a given cluster by the hypergeometric probability distribution[28]. For each cluster, the probability (p-value) of observing such an overlap by chance is calculated as:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}} \qquad (7)$$

where, G = the size of the genome; C = the number of genes in the genome having that attribute; n = the size of the query cluster; k = the number of genes in the cluster known to have that attribute[28].

Most of the p-values in Table 2 are quite small ($< 10^{-3}$) for the three clusters we are reporting here. These small p-values signify the relatively strong functional coherency of these clusters. How small must a p-value be in order for a cluster to be functionally coherent? FunSpec uses 0.01 as a cut off, which is arbitrary. For each of the clusters, we obtain p-values that are much smaller than 0.01, indicating the highly probable functional coherency of the clusters.

One of our goals in this paper is to test the validity of a reported interaction by using structural information about the interacting proteins in a cluster. Our idea is simple: first, find the structures of the two interacting proteins from the PDB[10]. If the experimental structure is not available in the PDB for any of the proteins, we predict its structure by comparative modeling. For comparative modeling, we used both CABS modeling[11] and I-TASSER[12-14]. However, the results shown here come only from using I-TASSER. Once, we have both structures, we dock them to predict the interaction complex. We can repeat this method to verify individual interaction in a cluster.

Here, we show an example of this approach. We find the homologs for the six proteins in cluster 14 shown in Figure 1. For the three proteins – YOL001W, YPL153C, and YGL058W – we retrieve the PDB structures having 100% identity as 2PK9 chain B, 1QU5 chain A, and 1AYZ chain A, respectively. For the other three proteins – YBR160W, YML032C, and YDL020C – the PDB homologs are 3EZR chain A (62% identity), 1KN0 chain A (53% identity), and 1A1I chain A (43% identity), respectively. For the latter three proteins, we predict their structures using the I-TASSER server [12-14]. I-TASSER reports the top five predictions for each submitted protein sequence, according to the c-score and the cluster size. We select the model that has the highest c-score out of the five returned models for each target sequence. I-TASSER also returns the top ten templates that it used for threading. We report the template that has the best sequence identity for the target protein sequence. For each unknown structure, Figure 3 shows the top prediction, the closest template, and the structural superposition of the predicted structure and the template. The c-scores for the models of YBR160W, YDL020C, and YML032C are 0.65, 0.41, and -0.54, respectively. We also compute the surface areas for each of the models and the reported template by using NACCESS which is an implementation of the methods described by Lee and Richards[29] and Hubbard, Campbell and Thornton[30]. The surface areas for the model for YBR160W and its template (PDB id 2PK9A) are 15,727Å^2 and 15,074Å^2, respectively which are similar. Also the surface areas of the model of YDL020C and its template (PDB id 1z1nx) are 33,655Å^2 and 33,482Å^2, respectively. The similarity in these surface areas can serve as a crude indication of the quality of the model returned from the server. In cluster 14, there are nine interactions. Four interactions involve YML032 whose model returned from the I-TASSER server is not a globular protein. This model is a very extended open structure. As a result, it would appear to have significant structural flexibility and thus not

be fully suitable for rigid body docking using Cluspro. We have performed docking for the other five interactions. Results of docking for these five interactions are shown in Figure 4. For each interaction, the figure shows the surface views of the docked complexes. To measure how strongly these docked complexes are bound, we have calculated the buried surface area for each docked complex. Table 3 shows the buried surface area and the ratio between buried surface area and total surface area of each of the docked complexes.

YOL001W has 100% sequence identity with 2PK9 chain B and the template used by I-TASSER to predict the structure of YBR160W is 2PK9 chain A. This suggests that there might be an interaction between YOL001W and YBR160W because of this known dimeric structure. We docked the homolog (2PK9 chain A) of YOL001W and the model for YBR160W. The docked complex, YBR160W.YOL001W, is shown in Figure 4(f). The ratio of the buried surface area to the total surface area for this complex is the largest among all the dimers as shown in Table 3. Therefore if we consider buried surface area relative to the total surface area of a complex as a measure of the strength of an interaction between two proteins, the complex YBR160W.YOL001W is expected to be more stable than the other dimers. This could also mean that this new interaction between YBR160W and YOL001W would be stronger than the other pair-wise interactions.

It is evident from Figure 4(a), (b), and (c) that protein YDL020C has at least two binding sites. YBR160W and YOL001W both bind to YDL020C at overlapping sites but YGL058W binds with YDL020C at a completely different binding site. Thus, the interactions YDL020C.YBR160W and YDL020C.YGL058W or YDL020C.YOL001W and YDL020C.YGL058W could occur simultaneously. Figure 1(d) shows the new core of cluster 14 with YML032C and its related interactions removed and the newly discovered interaction

YBR160W.YOL001W included. The docked complexes for these two set of mutually exclusive interactions, YGL058W.YDL020C.YBR160W and YGL058W.YDL020C.YOL001W, are shown in Figure 5(a) and (b) respectively. By analyzing the binding sites of YOL001W, we find that it has different binding sites to bind with YBR160W and YOL001W, thus making these two interactions concurrently possible. For a similar reason, the interactions YDL020C.YOL001W and YBR160W.YOL001W can occur simultaneously. The resultant trimers are shown in Figure 5(c) and (d), respectively. All other pair-wise interactions (d, e and f) in Figure 4 are mutually exclusive because of shared binding sites of the interacting proteins. Table 3 shows the list of all possible higher order complexes that can be modeled from the four protein molecules in the new core (shown in Figure 1(d)) of cluster 14. We also compute the buried surface areas of the trimers, as shown in Table 3. This table also shows that the ratio between the buried surface area and total surface area for the trimer YBR160W.YOL001W.YGL058W is bigger than that of the other trimer thus making the former more stable. For similar reason, we rank the tetramer YGL058W.YDL020C.YOL001W.YBR160W as more stable than the tetramer YGL058W.YDL020C.YBR160W.YDL020C.

5.4 Discussion

It is evident from the model for YML032C in Figure 3(c) that YML032C is a highly flexible protein. The results from disorder predictors[31] also show that this protein is disordered. High flexibility and disorder of this protein indicates that this could be a regulatory protein. Highly flexible and disordered proteins are functionally promiscuous as they can go through large and wide conformational changes while binding with other proteins or ligands[32]. Some disordered proteins attain tertiary structure of the binding site only when the binding with the ligand occurs. New methods that allow combining docking with folding of the disordered parts of a protein

structure have been recently proposed [33-39]. Flexible docking can predict protein-protein interaction complexes while allowing limited flexibility of the interacting proteins. Most methods consider ligand flexibility [37;38;39] and some address hinge motion, side chain flexibility, and docking with multiple conformations of a target protein obtained from multiple structures for the same protein in the PDB database [38]. However no docking algorithm can presently treat the high flexibility and disorder as found in YML032C.

We have used the ratio between the buried surface area and the total surface area of a protein complex as a measure for the strength of an interaction. Although we cannot definitely say whether an interaction actually happens or not from the value of this ratio, the value itself gives a certain level of confidence in that interaction.

5.5 Conclusion

This work has taken the approach of predicting new protein-protein interaction complexes and their interactions through docking of their molecular structures. Since not all complexes are available in the PDB, nor are they all likely to ever be available, we have relied upon comparative modeling and docking methods. Their recent improved reliability gives some justification for the use of these approaches. This methodology has the advantage that it can also identify interactions that could occur together or ones that are mutually exclusive. In addition indirect interactions through another intermediate protein can be identified. However, because of the lengthy computational times and the required human judgment to select models from the results of the prediction programs for comparative modeling and docking, this process cannot yet be fully automated. Nonetheless many such cases can be investigated, and it appears that the results can provide important new information.

In this computational prediction of interaction complexes, new interactions, and concurrency or exclusiveness of multiple interactions require two major computational steps – comparative modeling (I-TASSER server[14]) and docking (Cluspro server[22;23]). We plan to develop software that will use a cluster of protein interactions as input to produce final structures.

Validation of these predictions is an important task. At this time, we have not experimentally validated these predictions of new protein-protein interactions and their complexes. Because of the relatively few structures for protein complexes in the PDB database, we have not found clusters where the structures for the predicted complexes are available in the PDB database. Therefore, at this point, the correctness of our results depends on that of the underlying computational methods – techniques for comparative modeling, clustering, and buried surface area computations. Our theoretical predictions might be however useful for crystallographers to select targets for the X-ray crystallographic determination of protein complexes.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ARK, AK, and RLJ all contributed to the design, execution and writing of this work.

Acknowledgments

We acknowledge the assistance of Taner Z. Sen (Iowa State University) and Michael Zimmermann (L.H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University). We also thank S. Hubbard for generously providing his NACCESS program, which is an implementation of the methods described by Lee and Richards[29] and Hubbard, Campbell

and Thornton[30] to compute the buried surface areas of protein complexes. Grant sponsor: NIH; Grant numbers: R01GM073095, R01GM072014, and R01GM081680.

Bibliography

1. Young KH.: Yeast two-hybrid: so many interactions, (in) so little time.. Biol.Reprod. 1998 Feb;58(2):302-11.
2. Patra SM, Vishveshwara S.: Backbone cluster identification in proteins by a graph theoretical method. Biophys.Chem. 2000 Feb 14;84(1):13-25.
3. Sen TZ, Kloczkowski A, Jernigan RL.: Functional clustering of yeast proteins from the protein-protein interaction network. BMC.Bioinformatics. 2006;7:355.
4. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M.: BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535-D539.
5. Roberts MW, Ongkudon CM, Forde GM, Danquah MK.: Versatility of polymethacrylate monoliths for chromatographic purification of biomolecules. J.Sep.Sci. 2009 Jul 14.
6. Aloy P, Russell RB.: Ten thousand interactions for the molecular biologist. Nat.Biotechnol. 2004 Oct;22(10):1317-21.
7. von MC, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 2002 May 23;417(6887):399-403.
8. Aloy P, Russell RB.: Understanding and Predicting Protein Assemblies With 3D Structures. Comp Funct.Genomics 2003;4(4):410-5.
9. Anderson E, Demmel J, Bai Z, Bischof C, Blackford S, Dongarra J, Du Croz, Greenbaum A, Hammarling S, McKenney A, et al.: LAPACK Users' Guide 3rd Edition. Society for Industrial and Applied Mathematics 1999 . 1999.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE.: The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42.
11. Kolinski A.: Protein modeling and structure prediction with a reduced representation. Acta Biochim.Pol. 2004;51(2):349-71.
12. Zhang Y.: Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 2007;69(Suppl) 8:108-17.
13. Wu S, Skolnick J, Zhang Y.: Ab initio modeling of small proteins by iterative TASSER simulations. BMC.Biol. 2007;5:17.
14. Zhang Y.: I-TASSER server for protein 3D structure prediction. BMC.Bioinformatics. 2008;9:40.
15. Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T.: Automated server predictions in CASP. Proteins 2007;69(Suppl) 8:68-82.
16. Cozzetto D, Kryshtafovych A, Ceriani M, Tramontano A.: Assessment of predictions in the model quality assessment category. Proteins 2007;69(Suppl) 8:175-83.
17. Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T.: Assessment of CASP7 predictions for template-based modeling targets. Proteins 2007;69(Suppl) 8:38-56.

18. Comeau SR, Kozakov D, Brenke R, Shen Y, Beglov D, Vajda S.: ClusPro: performance in CAPRI rounds 6-11 and the new server. Proteins 2007 Dec 1;69(4):781-5.

19. Shen Y, Brenke R, Kozakov D, Comeau SR, Beglov D, Vajda S.: Docking with PIPER and refinement with SDU in rounds 6-11 of CAPRI. Proteins 2007 Dec 1;69(4):734-42.

20. Kozakov D, Brenke R, Comeau SR, Vajda S.: PIPER: an FFT-based protein docking program with pairwise potentials. Proteins 2006 Nov 1;65(2):392-406.

21. Comeau SR, Vajda S, Camacho CJ.: Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. Proteins 2005 Aug 1;60(2):239-44.

22. Comeau SR, Gatchell DW, Vajda S, Camacho CJ.: ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W96-W99.

23. Comeau SR, Gatchell DW, Vajda S, Camacho CJ.: ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics. 2004 Jan 1;20(1):45-50.

24. Vajda S, Kozakov D.: Convergence and combination of methods in protein-protein docking. Curr.Opin.Struct.Biol. 2009 Apr;19(2):164-70.

25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, et al.: Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet 2000 May; 25(1): 25-29

26. Bu D, Zhao Y, Cai L, Xue H, Zhu X, Lu H, Zhang J, Sun S, Ling L, Zhang N, et al.: Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Res. 2003 May 1;31(9):2443-50.

27. Robinson MD, Grigull J, Mohammad N, Hughes TR.: FunSpec: a web-based cluster interpreter for yeast. BMC.Bioinformatics. 2002 Nov 13;3:35.

28. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM.: Systematic determination of genetic network architecture. Nat.Genet. 1999 Jul;22(3):281-5.

29. Lee B, Richards FM.: The interpretation of protein structures: estimation of static accessibility. J.Mol.Biol. 1971 Feb 14;55(3):379-400.

30. Hubbard SJ, Campbell SF, Thornton JM.: Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J.Mol.Biol. 1991 Jul 20;220(2):507-30.

31. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B.: Improved disorder prediction by combination of orthogonal approaches. PLoS.One. 2009;4(2):e4433.

32. Nobeli I, Favia AD, Thornton JM.: Protein promiscuity and its implications for biotechnology. Nat.Biotechnol. 2009 Feb;27(2):157-67.

33. Coluzza I, Frenkel D.: Monte Carlo study of substrate-induced folding and refolding of lattice proteins. Biophys.J. 2007 Feb 15;92(4):1150-6.

34. Turjanski AG, Gutkind JS, Best RB, Hummer G.: Binding-induced folding of a natively unstructured transcription factor. PLoS.Comput.Biol. 2008 Apr;4(4):e1000060.

35. Verkhivker GM, Bouzida D, Gehlhaar DK, Rejto PA, Freer ST, Rose PW.: Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding. Proc.Natl.Acad.Sci.U.S.A 2003 Apr 29;100(9):5148-53.

36. Wright PE, Dyson HJ.: Linking folding and binding. Curr.Opin.Struct.Biol. 2009 Feb;19(1):31-8.

37. Jones G, Willett P, Glen RC, Leach AR, Taylor R.: Development and validation of a genetic algorithm for flexible docking. J.Mol.Biol. 1997 Apr 4;267(3):727-48.

38. Claussen H, Buning C, Rarey M, Lengauer T.: FlexE: efficient molecular docking considering protein structure variations. J.Mol.Biol. 2001 Apr 27;308(2):377-95.

39. Totrov M, Abagyan R.: Flexible ligand docking to multiple receptor conformations: a practical alternative. Curr.Opin.Struct.Biol. 2008 Apr;18(2):178-84.

Figures

Figure 1 - Examples of three clusters and their interactions with the nodes being the proteins and their names given

(a) cluster 10 (b) cluster 14 (c) cluster 15 (d) New core of cluster 14 after YML032C and YPL153C were removed is shown schematically with yellow being YBR160W, purple YGL058W, cyan YDL020C, and red YOL001W. All 6 edges of this tetrahedron correspond to pairs of proteins that interact, with the red edge being the newly proposed interaction.

Figure 2 - Method for structure prediction of a protein interaction pair

(a) Flowchart for obtaining each protein structure (b) Flowchart for docking two proteins to form a docked complex

Figure 3 - Comparative modeling for three unknown proteins in cluster 14 shown in Figure 2

a(1) Model for YDL020C a(2) One of the templates used by I-TASSER (PDB ID:1Z1NX) a(3) Superimposition of the model and the template (RMSD = 0.410) b(1) Model for YBR160W b(2) One of the templates used by I-TASSER (PDB ID:2PK9A) b(3) Superimposition of the model and the template (RMSD = 0.77) c(1) Model for YML032C c(2) Template used by I-TASSER (PDB ID:1WORA) c(3) Superimposition of the model and the template (RMSD = 0). The difference in buried surface area for the model in a(1) and template in a(2) is 173 Å2 and that is in b(1) and b(2) 654 Å2.

Figure 4 - Models built for the interactions in the core of cluster 14

Buried surface areas of the dimers (a) YDL020C.YBR160W(5,603Å2) (b) YDL020C.YOL001W (4,517Å2) (c)YDL020C.YGL058W(4,295Å2) (d)YBR160W.YGL058W(3,408Å2) (e)YOL001W.YGL058W(2,162 Å2) (f)YBR160W.YOL001W (5,779 Å2).

Figure 5 - Trimers built from pairs of interactions of proteins in the core of cluster 14.

Buried surface areas of the trimers (a) 9,898 Å2 for YGL058W.YDL020C.YBR160W (the docked complex if interactions a and c in Figure 4 occur simultaneously) (b) 8,812 Å2 for YGL058W.YDL020C.YOL001W (the docked complex if interactions b and c in Figure 4 occur simultaneously) (c) 7,941 Å2 for YBR160W.YOL001W.YGL058W (the docked complex if interactions e and f in Figure 4 occur simultaneously) (d) 10,296 Å2 for YDL020C.YOL001W. YBR160W (the docked complex if interactions b and f in Figure 4 occur simultaneously).

Tables

TABLE 1. Functions of proteins in clusters 10, 14, 15 of yeast protein network-2.0.41

Protein name	Function	Function type
	Cluster 10	
YBR160W	Catalytic subunit of the main cell cycle cyclin-dependent kinase	***Cell cycle***
YGL058W	Ubiquitin-conjugating enzyme (E2), involved in postreplication repair (with Rad18p), sporulation, telomere silencing, and ubiquitin-mediated N-end rule protein degradation (with Ubr1p)	Protein repair/ degradation
YLR200W	Subunit of the heterohexameric Gim/prefoldin protein complex involved in the folding of alpha-tubulin, beta-tubulin, and actin	Protein folding
YOL001W	Cyclin, negatively regulates phosphate metabolism	***Cell cycle***
YPR119W	B-type cyclin involved in cell cycle progression	***Cell cycle***
	(b) Cluster 14	
YBR160W	Catalytic subunit of the main cell cycle cyclin-dependent kinase	***Cell cycle***
YOL001W	Cyclin, negatively regulates phosphate metabolism	***Cell cycle***
YPL153C	Protein kinase, required for cell-cycle arrest in response to DNA damage	***Cell cycle***
YML032C	Stimulates strand exchange by facilitating Rad51p binding to single-stranded DNA	***Cell cycle***
YDL020C	Transcription factor that stimulates expression of proteasome genes Type	Protein degradation
YGL058W	Ubiquitin-conjugating enzyme (E2), involved in postreplication repair (with Rad18p), sporulation, telomere silencing, and ubiquitin-mediated N-end rule protein degradation (with Ubr1p)	Protein repair/ degradation
	(c) Cluster 15	
YGL058W	Ubiquitin-conjugating enzyme (E2), involved in postreplication repair (with Rad18p), sporulation, telomere silencing, and ubiquitin-mediated N-end rule protein degradation (with Ubr1p)	Protein repair/ degradation
YBR160W	Catalytic subunit of the main cell cycle cyclin-dependent kinase	***Cell cycle***
YEL003W	Subunit of the heterohexameric cochaperone prefolding complex which binds specifically to cytosolic chaperonin and transfers target proteins to it	Protein folding
YDL020C	Transcription factor that stimulates expression of proteasome genes Type	Protein degradation
YHR200W	Non-ATPase base subunit of the 19S regulatory particle (RP) of the 26S proteasome	Protein degradation
YPL153C	Protein kinase, required for cell-cycle arrest in response to DNA damage	***Cell cycle***

TABLE 2. MIPS functional classification and GO(Gene Ontology) assignments of biological processes and molecular functions for clusters 10, 14, and 15

Cluster #	# proteins	GO molecular function	GO biological process	MIPS functional classification
10	5	Cyclin-dependent protein kinase regulatory activity (5×10^{-5}) Tubulin binding (4×10^{-3})	Regulation of cyclin-dependent protein kinase activity (6×10^{-5}) Negative regulation of phosphate metabolic process (9×10^{-4})	Enzymatic activity regulation / enzyme Regulator (5×10^{-4}) Regulation of phosphate metabolism(9×10^{-3})
14	6	Recombinase activity (2×10^{-3}) DNA strand annealing activity (3×10^{-3})	Postreplication repair (1×10^{-4}) regulation of cell cycle (5×10^{-4}) Response to DNA damage stimulus (7×10^{-4})	DNA repair (3×10^{-4}) G2/M transition of mitotic cell cycle (7×10^{-4})
15	6	Protein serine/threonine/tyrosine kinase activity (5×10^{-3})	Regulation of cell cycle (6×10^{-4}) Negative regulation of meiotic cell cycle (10×10^{-4})	Proteasomal degradation (ubiquitin/proteasomal pathway) (2×10^{-4})

Table 3. Buried surface area (SA) of the docked complexes
(the order of the complexes in this table is the same as in Figure 4(a–f) and Figure 5(a–d) for dimers and trimers, respectively)

Interacting complex	Buried SA(Å^2)	2*Buried SA/(Total SA)
Dimers		
YDL020C : YBR160W	5,603	0.23
YDL020C : YOL001W	4,517	0.20
YDL020C : YGL058W	4,295	0.20
YBR160W : YGL058W	3,408	0.29
YOL001W : YGL058W	2,162	0.21
YBR160W:YOL001W	5,779	0.41
Trimers		
YGL058W.YDL020C.YBR160W	9,898	0.34
YGL058W.YDL020C.YOL001W	8,812	0.33
YBR160W.YOL001W.YGL058W	7,941	0.44
YDL020C.YOL001W.YBR160W	10,296	0.33
Tetramers		
YGL058W.YDL020C.YBR160W.YDL020C	15,501	0.34
YGL058W.YDL020C.YOL001W.YBR160W	14,591	0.42

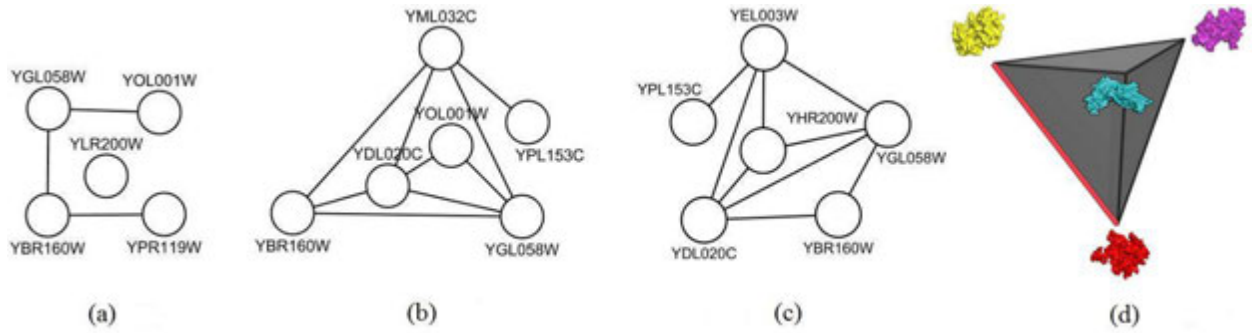


Figure 32.

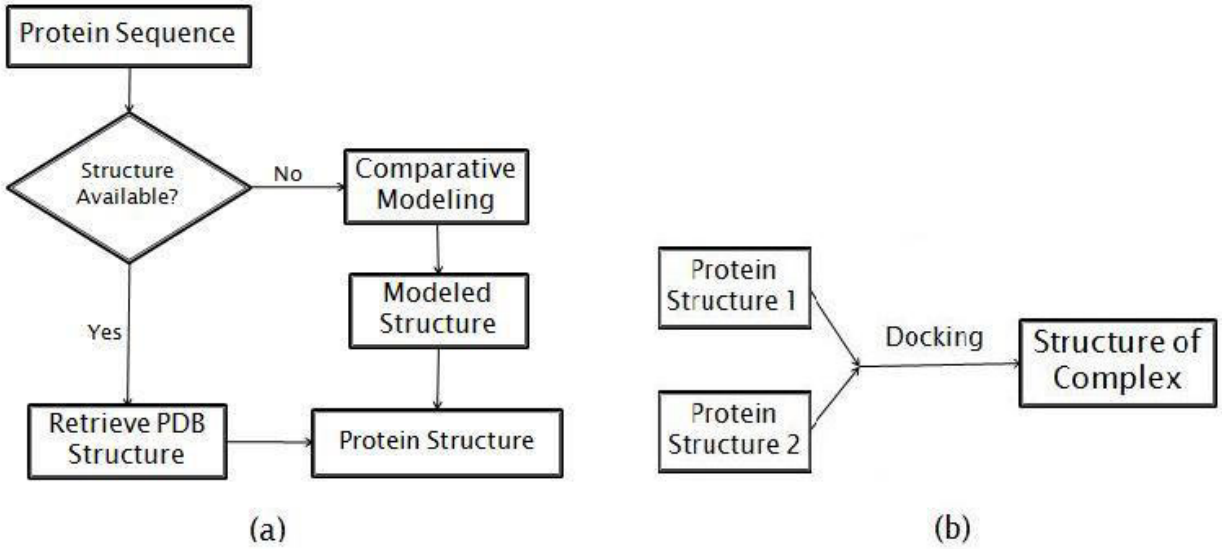


Figure 33.

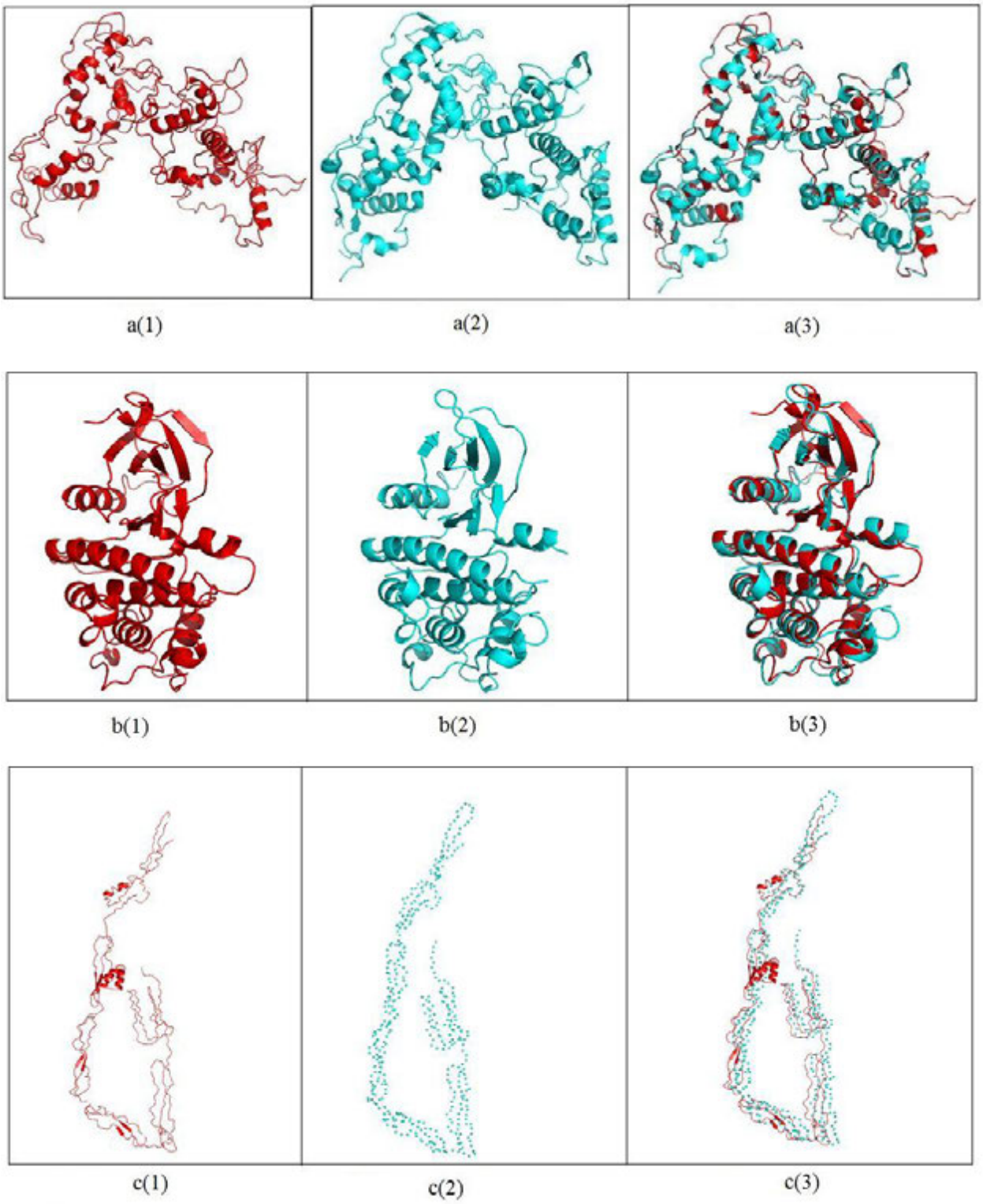


Figure 34.

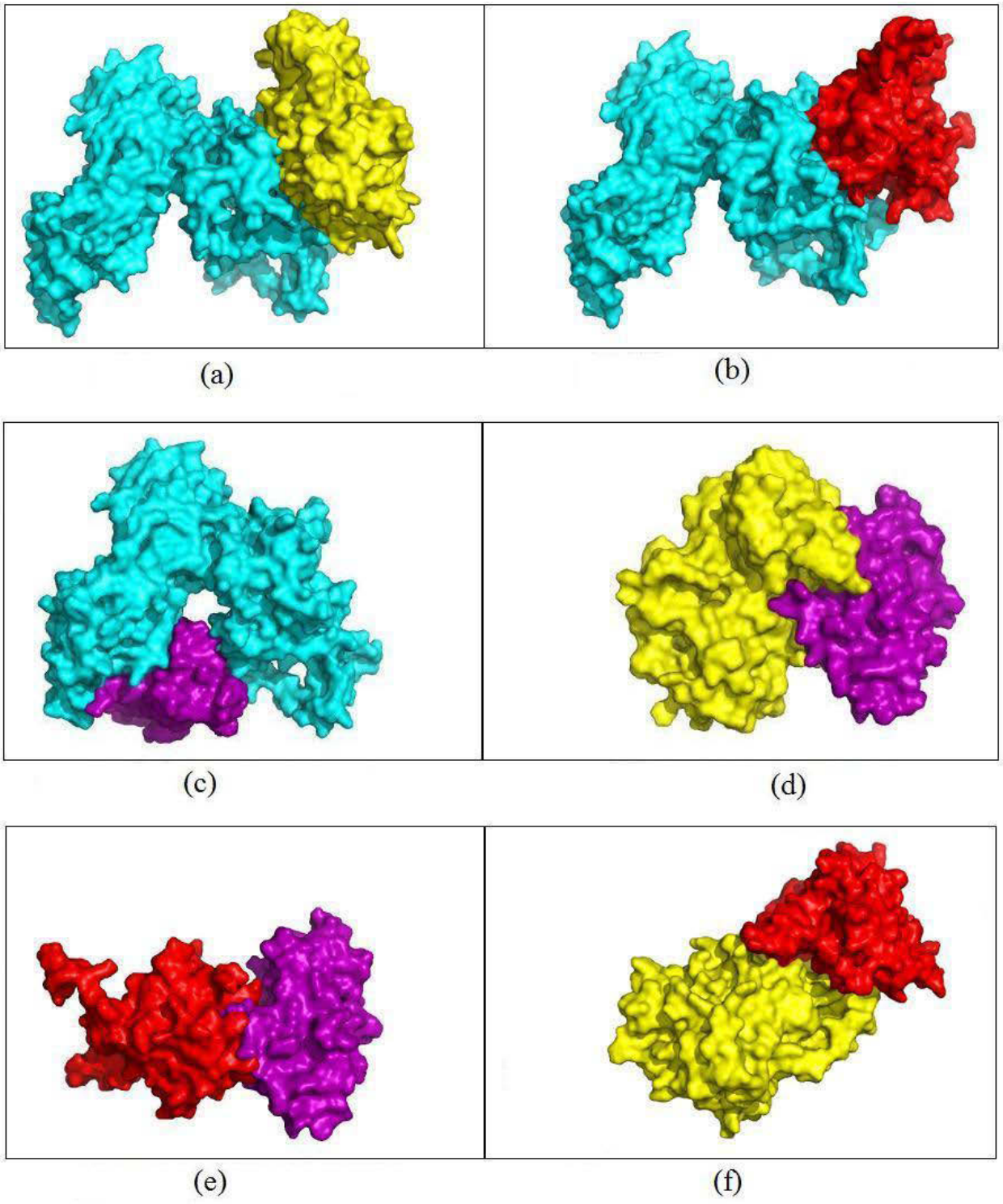


Figure 35.

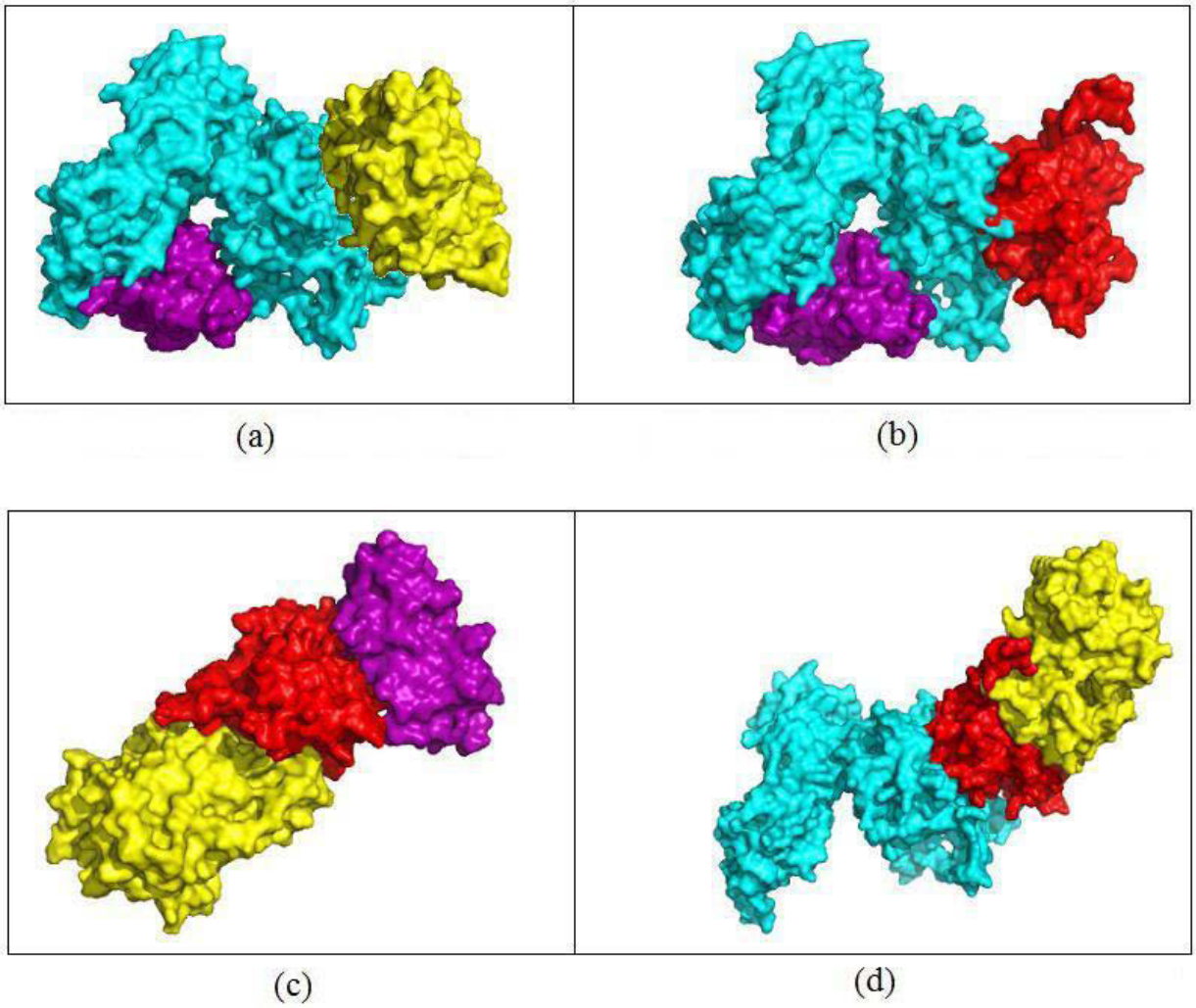


Figure 36.

CHAPTER 6. CONCLUSIONS

FBA and TIM are both beta barrel proteins that usually function as dimers in most organisms. Each subunit of the structure has a central barrel comprised of eight β-strands surrounded by eight α-helices. There are loops running from a β-strand to the subsequent α-helix and from an α-helix to the subsequent β-strand connecting all the strands and helices, except for the terminal secondary structure elements. They have similar architectures for their cores formed by the strands, with an RMSD of 4.8 Å. Despite the 3D fold similarity of these two proteins, they have low sequence identity. However the functional residues and the catalytic microenvironment in each structure are highly conserved across the species.

Each subunit has a catalytic pocket. In FBA, residues from the partner subunits contribute to the formation of the catalytic pocket. In TIM, however, the catalytic pocket in each subunit is structurally independent of the partner subunit, with oligomerization pushing the catalytic lysine on interface loop 1 towards the center of the catalytic pocket and positioning it into a catalytically favorable position.

6.1 Summary of the Results and their Application to the Research Problem

Dimerization is required for both enzymes to achieve their catalytic capability. We have investigated the dynamics of these two enzymes in different oligomeric states, dimer and tetramer. By applying ENM to model the dynamics of the FBA structure, we find that oligomerization stabilizes the structural components that form the interface. Consequently, this stabilizes the residues in the interface region of the partner subunit that take part in forming the catalytic environment. Also oligomerization facilitates the specific motions of the loops and the residues on those loops achieve the required level of fluctuation to aid the catalysis. Application

of ENM to model the dynamics of the TIM structure shows that oligomerization attenuates the interface mobility and this attenuation of fluctuation along the interface region in turn increases the fluctuations across the functional loops 6, 7, and 8.

We find that the FBA and TIM functional loops are strongly synchronized within each structure. This synchronization of the functional loops with one another together with their appropriate dynamics may help these structures achieve their high rates of activity. We also find that the functional loops in FBA are well coordinated with the functional loops in TIM. This is a consequence of the high level of structural similarity between the cores of the two proteins.

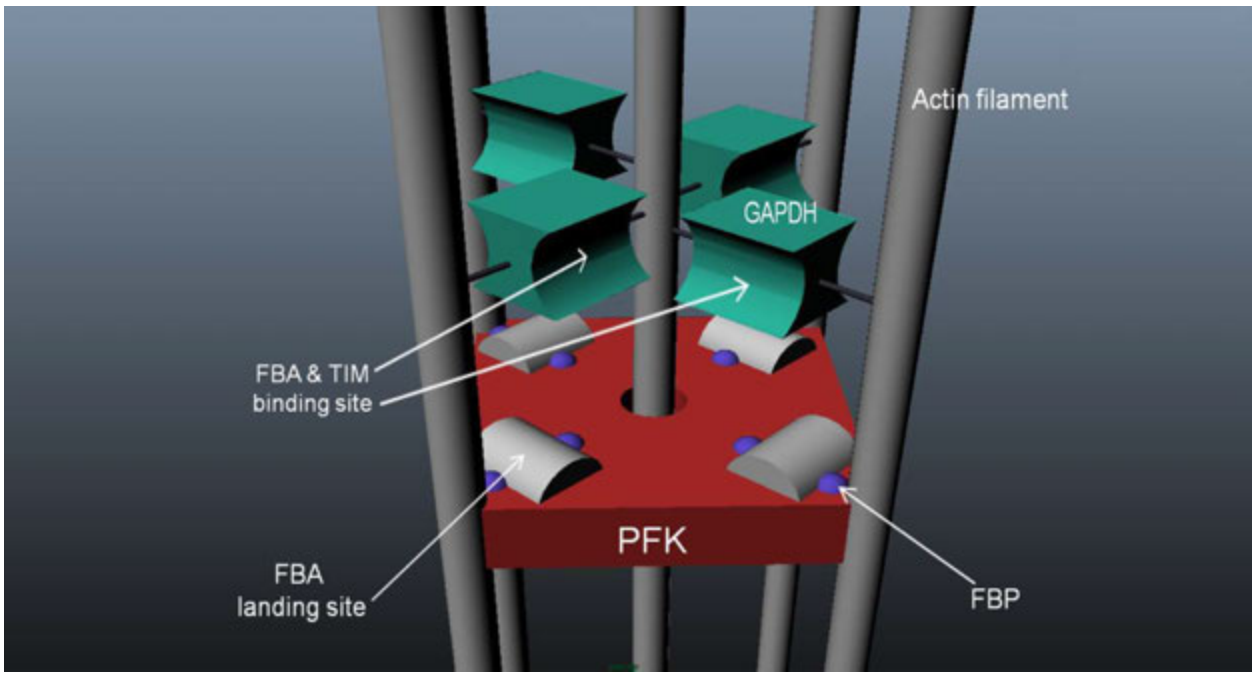


Figure 37. A conceptual model for the localization of the FBA and TIM enzymes between the neighboring proteins in the glycolysis cycle and actin fibers, which may facilitate the substrate transfer from FBA to TIM. The acronyms are those defined in Fig. 1 of Chapter 1. The actin filaments and bundles are known to be attached to the glycolytic enzymes GAPDH and PFK, and all these components act together to confine FBA and TIM within some limited space.

On the other hand, analysis of the sequence and structure of FBA and TIM identify a 'phosphate gripper' motif on one of the functional loops on each protein – FBA conserves this

motif on the N-terminus of loop 6 and TIM has it located at the tip of its most mobile functional loop 6.

FBA and TIM are two enzymes that function between two larger size enzymes – PFK and GAPDH. The actin filaments and bundles interact with these larger enzymes PFK and GAPDH [6-10] in such a way that FBA and TIM function within a localized environment. Figure 1 shows a conceptual depiction of the localized environment where this may happen.

Considering the localization of environment where FBA and TIM operate and from our findings about the coordination of the functional dynamics of these two enzymes, we propose the following hypotheses regarding how of FBA and TIM function as a joint molecular machine.

I. FBA hunts its substrate FBP by using its tentacle-like 'phosphate gripper' region on loop 6. The motions of the functional loops help the substrate to move into the catalytic pocket between loop 5 and loop 6 where the cleaving of the substrate occurs to yield its products GAP and DHAP. The contact between the tip of loop 6 and the coil between helix 8A and helix 8B of the partner subunit may exert a force to push loop 6 outward and to eject the products from the catalytic pocket. The DHAP is then caught by the 'phosphate gripper' of the front loop 6 of TIM when both enzymes come into the proper position.

II. An alternative hypothesis for the DHAP transfer from FBA to TIM is that a part of the 'phosphate gripper' remains attached to the substrate before and the product after catalysis. After the product is formed in the FBA catalytic pocket, loop 6 opens outwards still carrying the DHAP product and reaches out towards the catalytic site of TIM where the similar 'phosphate gripper' region of loop 6 of TIM replaces the phosphate gripper of FBA.

III. Another alternative hypothesis for the substrate transfer is that the ‘phosphate grippers’ in both structures are disordered and open in both structures. The disordered ‘phosphate gripper’ in open FBA and TIM structure is conducive to reach further away from the catalytic site and attach to the substrate thus facilitating its recruitment into the cavity; this would be a process akin to the flycasting mechanism for disordered parts of proteins that was proposed by Onuchic and others [11-13]. Once the substrate is brought into the catalytic pocket of FBA and catalysis is completed, then loop 6 swings out with the substrate still bound to the loop that becomes sufficiently disordered to reach towards the phosphate gripper of TIM or even towards the catalytic pocket of TIM.

The internal synchronized dynamics of the functional loops of FBA and TIM may correlate with their enzymatic rates and the dynamics of the functional loops between these enzymes may also be synchronized. Thus FBA and TIM can operate as a synchronized conjugate machine. This FBA-TIM conjugate machine takes its substrate FBP through the ‘phosphate gripper’ of loop 6 of FBA and releases its product GAP into the opening of the TIM catalytic pocket.

Table 1. Oligomeric states and PDB homologs of the enzymes at four steps (3, 4, 5, and 6) of the glycolysis pathway shown in Figure 1 of Chapter 1.			
Step #	**Protein Name (Standard:Systematic)**	**Oligomeric State and # of Residues (Yeast)**	**PDB Code for a Homolog**
3	PFK PFK1:YGR240C PFK2:YMR205C PFK1:PFK2	PFK1-homo tetramer:987/chain PFK2-homo tetramer:959/chain PFK1:PFK2 complex-hetero dimer	3080:Sequence Id 98% with Yeast Alpha Subunit: A,C,E,G Beta Subunit: B,D,F,H
4	FBA:YKL060	Dimer:359/chain	1ZEN , 1B57 from bacteria; sequence identity with yeast 48%
5	TIM:YDR050C	Dimer:248/chain	1YPI, 7TIM from yeast
6	GAPDH TDH1:YJL052W, TDH2:YJR009C, TDH3:YGR192C	Tetramer: 332/chain	TDH1:3PYM with 100% sequence Identity
Note: Standard and Systematic protein names are taken from www.yeastgenome.org site			

Such FBA-TIM protein machinery provides a mechanism to upload the PFK product and after the necessary conversion delivers it to GAPDH as the latter one's substrate, a process described in the schematic diagram of Figure 1(B) in Chapter 4. This brings us one step closer to the construction of a protein machine of PFK-(FBA-TIM)-GAPDH. Table 1 shows the names of these four glycolytic proteins in *S.cerevisiae*, their oligomeric states, the number of residues, and the PDB structures for their homologs. It is noteworthy that PFK is a hetero-dimer – a PFK1:PFK2 conjugate complex. Both PFK1 and PFK 2 are homo-tetramers. The PFK complex has four pairs of catalytic sites. Each pair of sites consists of two catalytic sites – one site coming from PFK 1 and the other coming from PFK 2. The distance between the sites of a pair is 40Å which is about the same distance between the two catalytic sites of an FBA dimer. This is somewhat suggestive that each pair of the PFK catalytic sites might be a binding site for an FBA structure. On the other hand, functional GAPDH is a tetrameric structure. Observation and

analysis of the structure indicate that it has two pairs of catalytic sites – where each pair of sites is a binding location for either an FBA or a TIM structure. So a machine of FBA-GAPDH-TIM can transfer the products from FBA and TIM to GAPDH. Moreover these enzymes function in a compartment formed by the network of actin filament. Their proximity is mediated by the reaction between actin and the larger size enzymes such as PFK and GAPDH. Within this constrained environment FBA-TIM conjugate machine shuttles products from PFK to GAPDH. Using this structural and biochemical information, the mechanism of product transfer from PFK to GAPDH can be depicted with a structural model in Fig. 2. In step 1, FBA binds at one pair of the catalytic sites of PFK and loads its substrate from the PFK products. In step 2, an FBA-TIM machinery is formed and the product FBA is transferred to TIM as its substrate. In step 3, both FBA and TIM bind with GAPDH which loads its substrates from both of the former enzymes.

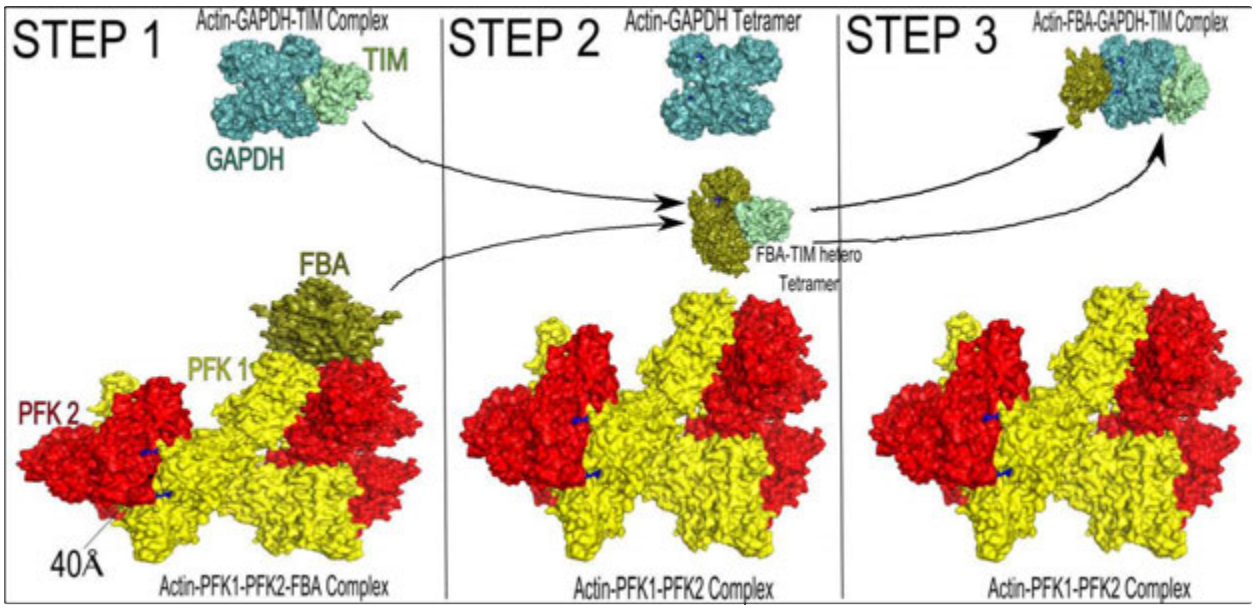


Figure 38. A structural model using the PDB structures for the steps shown in Fig. 2 of Chapter 1. Step 1 – FBA binds at one pair of the catalytic sites of PFK and loads its substrate from the PFK products; Step 2 – FBA-TIM machinery is formed and substrate transfer from FBA to TIM occurs; Step 3 – Both FBA and TIM bind with GAPDH and loads its substrate from both.

6.2 Future Research

While working on the mechanism of FBA-TIM interaction, I have also engaged in several other projects that are based on the experience I gathered from this research.

One protein machine built from enzymes for successive reactions along a metabolic pathway and whose assembly is known is the Fatty Acid Synthase machine, and in fact there are two very different structures, from different organisms. In principle, it could serve as a testbed for the assembly process and mechanism for transfer of substrate that we have been attempting to construct in our project for the glycolysis cycle. However, it presents a case with an unexpectedly high level of complexity, which explicitly points up why assembly by using only the simple concept of proximity between subsequent enzymes may be an insufficient principle for guiding the assembly of pathway-related enzymes into large machines.

ACP Movement Pathway in Fatty Acid Synthase Structure

Fatty Acids are synthesized in the Fatty Acid Synthesis Pathway from acetyl-CoA and malonyl-CoA precursors. There are two types of Fatty Acid Synthases: FAS I (Type I) and FAS II (Type II). FAS II is generally found in prokaryotes, plants, fungi, and parasites, as well as in mitochondria. FAS II is in general a set of separate enzymes. However, FAS I are generally found mammals, fungi, and yeasts where all catalytic domains are integrated into one large structure, where all synthesis steps occur in the separate enzyme domains. Mammalian FAS I is a large dimeric structure, whereas fungal and yeast FAS I is an even larger dodecameric protein.

Each subunit of the functionally active homo-dimeric mammalian FAS I structure has seven domains that are required for fatty acid synthesis – MAT, KS, KR, DH, ER, ACP, and TE. The Figure 4 shows the domains and substrate flow in such a structure except the ACP and TE

domains because of the absence of their structural information in the corresponding PDB structure (PDB Ids 2VZ9 and 2VZ8). The residue range for different domains and linker regions – KS: 1 ~ 406; KS-MAT Linker: 407 ~ 427; MAT: 428 ~ 815; MAT-DH Linker: 816 ~ 857; DH: 858 ~ 969; DH-Core Linker and Core and Core-ER Linker: 970 ~ 1629; ER: 1630 ~ 1850; ER-KR Linker: 1851 ~ 1869; KR: 1870 ~ 2100;

MAT – Malonyl/Acetyl Transferase: This enzyme transfers the acetyl group from acetyl-CoA to the ACP (Acyl Carrier Protein). This transformation produces acetyl-ACP complex which is carried to the KS domain for condensation with malonyl group of malonyl-CoA. Similar to acetyl-ACP, malonyl group of malonyl-CoA is also transferred to the ACP domain to form malonyl-ACP complex which gets transported to the KS domain to facilitate the condensation process as described below.

KS – Ketoacyl Synthase (or β-Ketoacyl-ACP Synthase): Condensation and Decarboxylation take place in this domain. Acetyl group of acetyl-ACP (CH_3CO-S-ACP) transfers to the KS and forms acetyl-KS complex (CH_3CO-S-KS) which reacts with the malonyl-ACP (COO-CH_2CO-S-ACP) to release CO_2 and form a condensed complex, aceto-acetyl-ACP (CH_3CO- CH_2CO-S-ACP). This complex is transferred to the KR domain.

KR – Ketoacyl Reductase (or β-Ketoacyl-ACP Reductase): Reduction of aceto-acetyl-ACP (a β-Ketoacyl-ACP) is catalyzed by KR and forms D-B-hydroxy butryl-ACP ($CH_3CH(OH)CH_2CO$-S-ACP), a β-hydroxy-acyl-ACP. This product is transported to the DH domain for a β-carbon modification.

DH – Dehydratase (or β-hydroxy-aceto-ACP Dehydrogenase): In the DH domain, D-B-hydroxy butryl-ACP is dehydrogenated by losing a water molecule and forming a carbon-carbon

double bond. The new complex is β-Enoyl-ACP. In the first of cycle, this complex is names as crotanyl-ACP (CH_3CH=$CHCO$-S-ACP). This goes through further reduction in the ER domain.

ER – Enoyl Reductase (or 2,3-trans Enoyl-ACP Reductase or β-Enoyl-ACP Reductase): Similar to the KR domain, an NADPH is oxidized to NADP in this domain and β-Enoyl-ACP is reduced to Acyl-ACP which, in the first cycle, is butyryl-ACP ($CH_3CH_2CH_2CO$-S-ACP).

This butyryl-ACP complex is transferred back to KS and the process is repeated for six more cycles and a 16-carbon β-Enoyl-ACP, palmitate-ACP complex is produced. Subsequently, the 16-carbon product, palmitate is released from ACP as free fatty acids by a Thioesterase (TE) domain.

Fungal dodecameric FAS I structure has over 20 thousand residues. The structure consists of a wheel and two domes where the wheel sits between the two domes. Figure 5 shows the structural components of this machine.

The activation, priming, elongation, and termination of fungal FAS are carried out by seven domains: (1) Phosphopantetheine Transferase (PPT), (2) Acyl Transferase (AT), (3) Malonyl/palmitoyl Transferase (MPT), (4) Ketoacyl Reductase (KR), (5) Dehydratase (DH), and (6) Enoyl Reductase (ER). Acyl Carrier Protein is an important flexible part of this structure. It is activated by binding with Phosphopantethein catalyzed by the Phosphopantethein Transferase (PPT) domain of the synthase. The elongating fatty acid is attached at one end of the phosphophantetheine (PP) while the other end is attached to the ACP as shown in Figure 6.

ACP is believed to have the dynamics to move from one active site to the next active site whereas stretching of its PP arm helps it bring the elongating fatty acid to the respective catalytic domain.

Previously, Simon Jenni *et al.* proposed that each ACP in fungal FAS interacts with a unique and distinct set of active sites, based on their identification of a comprehensive sets of active sites closest to the anchor points of each ACP [3]. Using an ENM simulation of FAS dynamics, we completed a similar analysis by finding the set of active sites which come closest to each ACP at any point in the first twenty modes of motion.

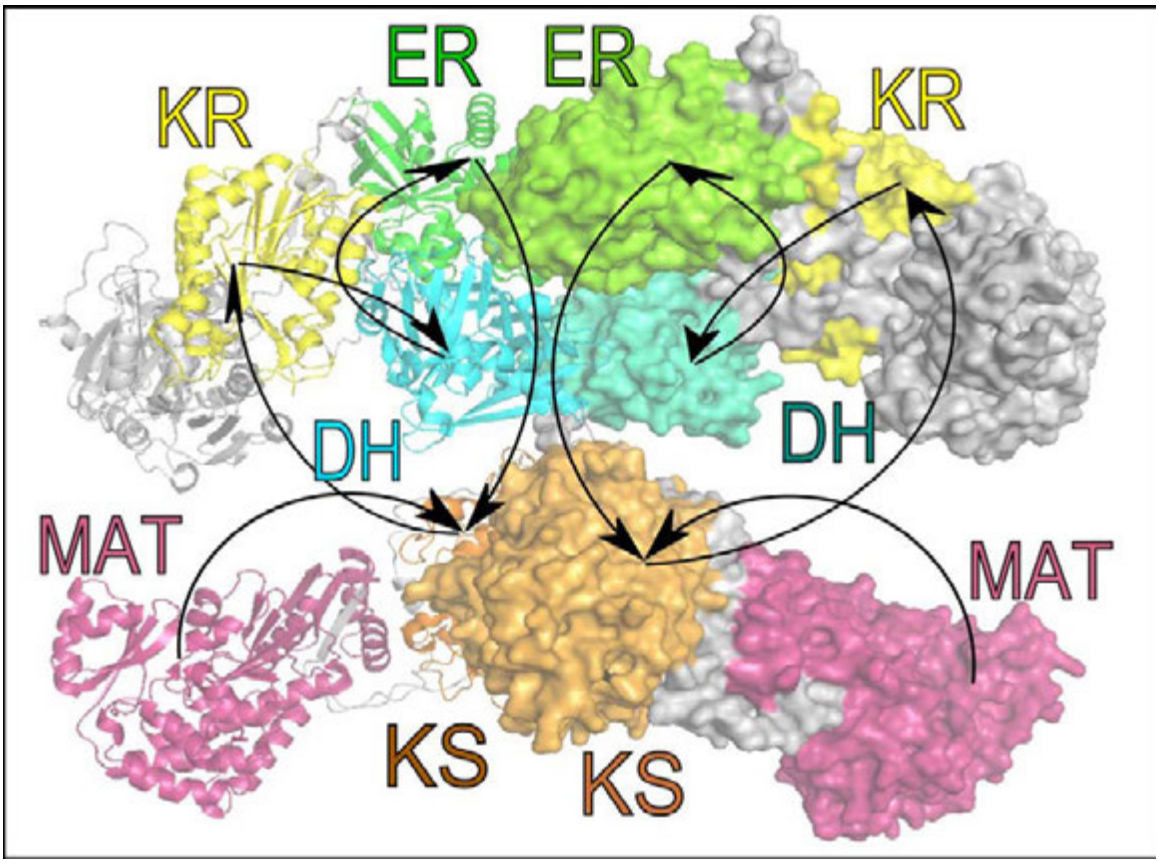


Figure 39. The enzymatic domains and linker regions of a mammalian fatty acid synthase based on PDB Id 2VZ9. Non enzymatic domains and domain connecting linker regions are shown in gray colors. Arrows indicate the direction of catalytic steps from one domain to another. By following the direction of the arrows, the elongation cycle, KS → KR → DH → ER → KS, can be traced. The final step for the release of palmitic acid from ACP catalyzed by the TE domain and the mobile ACP domain are not shown because of poor resolution in those parts.

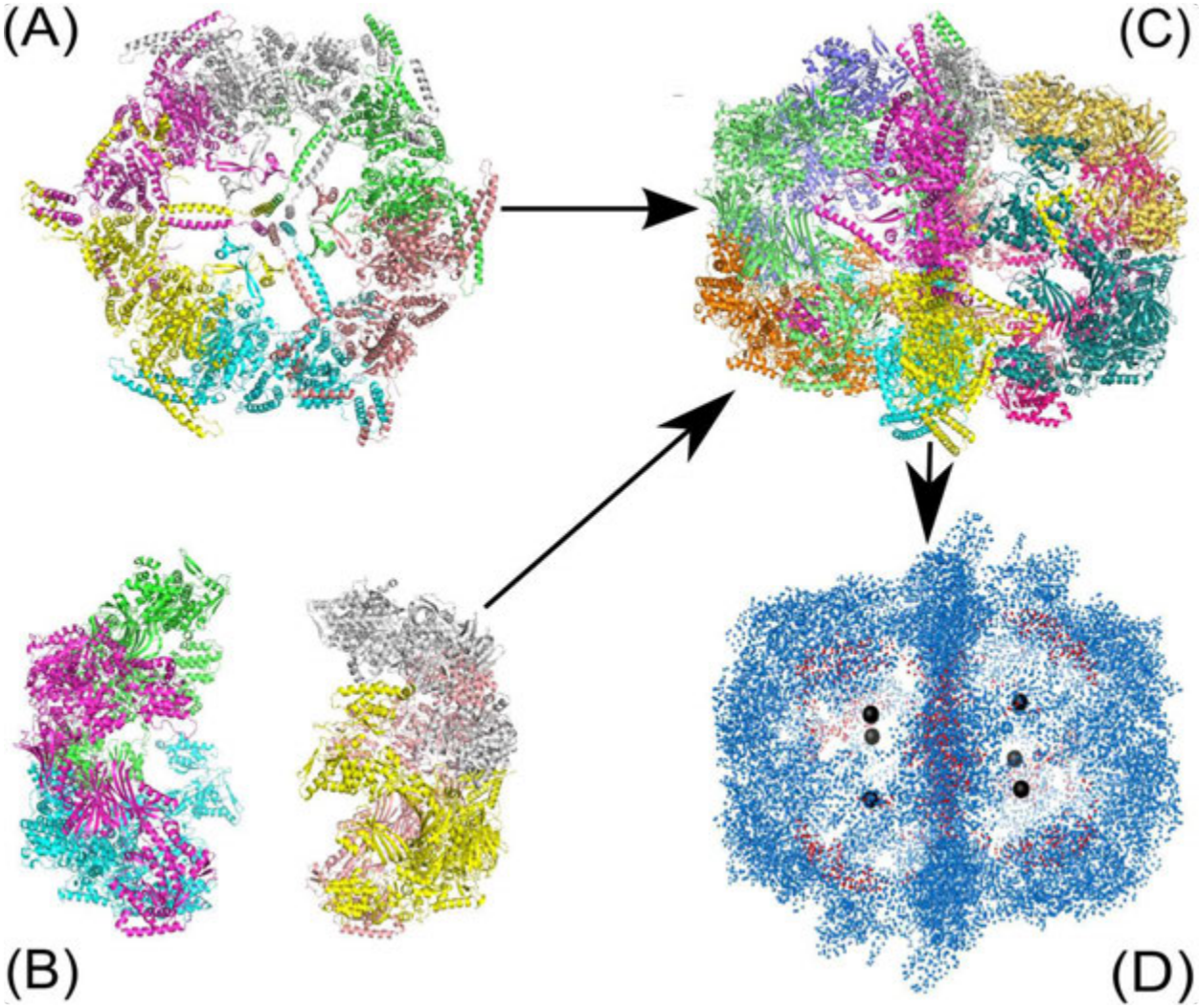


Figure 40. Structure of a Fungal Fatty Acid Synthase. (A) Wheel (PDB id 2UVB) (B) Two domes (2UVA) (C) Wheel sits between domes (D) A coarse grain model of the macromolecule – each black sphere indicates the center of each chamber. Different chains are shown in different colors.

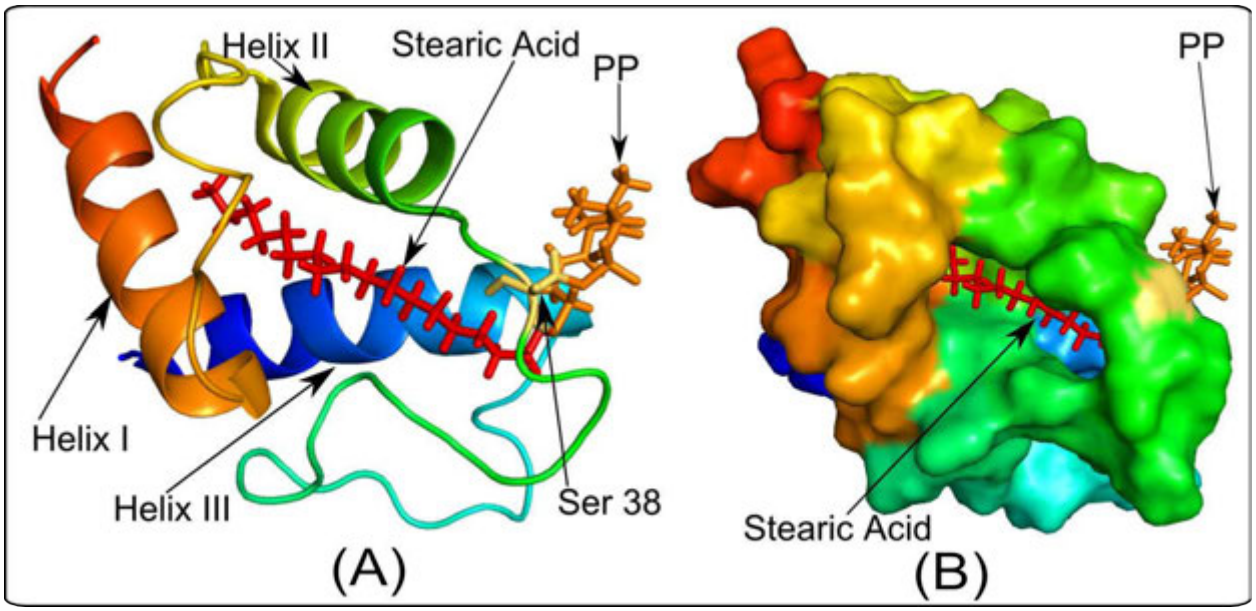


Figure 41. (A) Cartoon view of ACP which is bound with PP and elongating fatty acid – Ser 38 on ACP is the binding site for PP (B) Surface view of the same structure in (A) shows how the fatty acid is rested in the cavity formed inside ACP. Figure generated based on the structure with PDB Id 2FVA.

Bibliography

[1] R. L. Marsden, L. J. McGuffin, and D. T. Jones, "Rapid protein domain assignment from amino acid sequence using predicted secondary structure," *Protein Sci.*, vol. 11, no. 12, pp. 2814-2824, Dec. 2002.

[2] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. Database issue, p. D535-D539, Jan. 2006.

[3] S. Jenni, M. Leibundgut, D. Boehringer, C. Frick, B. Mikolasek, and N. Ban, "Structure of fungal fatty acid synthase and implications for iterative substrate shuttling," *Science*, vol. 316, no. 5822, pp. 254-261, Apr. 2007.

[4] "Links to Metabolics and Pathway Resources: http://www.bmrb.wisc.edu/metabolomics/,", Accessed 16 Dec. 2012.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, Jan. 2000.

[6] G. M. Cooper, "A Molecular Approach, 2nd Edition", Accessed online on 2/19/2013: http://www.ncbi.nlm.nih.gov/books/NBK9908/, http://www.ncbi.nlm.nih.gov/books/NBK9908/.

[7] S. L. Lowe, C. Adrian, I. V. Ouporov, F. Weingeh, and K. A. Thomasson, "Brownian dynamics simulations of glycolytic enzyme subsets with F-actin," *Biopolymers*, vol. 70, no. 4, pp. 456-470, Dec. 2003.

[8] F. Weingeh, S. L. Lowe, and K. A. Thomasson, "Brownian dynamics of interactions between glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mutants and F-actin," *Biopolymers*, vol. 73, no. 5, pp. 533-541, Apr. 2004.

[9] I. V. Ouporov, H.R. Knull, and K. A. Thomasson, "Brownian dynamics simulations of interactions between aldolase and G- or F-actin," *Biopphys J*, vol. 76, no. 1, pp. 17-27, Jan. 1999.

[10] N. Y. Forlemu, E.N. Njabon, K. L. Carlson, E.S. Schmidt, F. Weingeh, and K. A. Thomasson, "Ionic strength dependence of F-actin and glycolytic enzyme associations: a Brownian dynamics simulations approach," *Proteins*, vol. 79, no. 10, pp. 2813-2927, Aug. 2011.

[11] B.A. Shoemaker, J.J. Portman, P.G. Wolynes, "Speeding molecular recognition by using the folding funnel: the fly-casting mechanism," *PNAS*, vol. 97, no. 16, pp. 8868-8873, Aug. 2000.

[12] Y. Levy, J.N. Onuchic, P.G. Wolynes, "Fly-casting in protein-DNA binding: frustation between protein folding and electrostatics facilitates target recognition," *JACS*, vol. 129, no. 4, pp. 738-739, Jan. 2007.

[13] E. Trizac, Y. Levy, P.G. Wolynes, "Capillarity theory for the fly-casting mechanism," *PNAS*, vol. 107, no. 7, pp. 2746-2750, Aug. 2010.

APPENDIX A. COMPUTATIONAL TESTING OF PROTEIN-PROTEIN INTERACTIONS

Modified from a paper published in the peer reviewed conference, *IEEE BIBM 2009*

Ataur R. Katebi, Andrzej Kloczkowski, Robert L. Jernigan

Abstract

Abundant protein interaction data are currently available. These interaction networks are important sources of information about how biological systems function. In the present work the yeast protein interaction network is clustered and the individual clusters are investigated for functional relationships among the member proteins. 3D structural models of the proteins in a cluster have been built. We investigate the docking of proteins within the cluster, verifying reported and predicting some unreported interactions.

A.1 Introduction

Because of the use of high throughput experimental methods such as yeast two-hybrid screening[1], the number of reported protein-protein interactions has increased dramatically. To extract meaningful information from this interaction data set, clustering of the interacting proteins is an established method. Sen *et al*.[2] used an eigenmode analysis clustering method to cluster the interacting proteins with a spectral clustering method. Patra *et al.* [3] have shown that functionally significant clusters can be extracted from the dominant eigenvalues of a modified contact matrix known as the Kirchhoff matrix.

The BioGrid database has published different versions of the yeast protein interaction data with incremental numbers of proteins and their interactions[4]. Some limited attempts have been

made to construct spatial interaction clusters from these datasets. The results show that such clusters have functional relationships, which can then be used to predict undiscovered interactions among proteins in the same cluster[2]. However, the protein interaction data obtained from the high-throughput screening methods such as the yeast two-hybrid method[1] and affinity purification techniques[5] are highly error prone. Approximately, 30–60% false positives and 40–80% false negatives have been estimated for these methods [6;7]. Therefore, predicting new interactions or drawing any conclusions from this interaction dataset requires some reliable validation of the interactions. Other complementary source of information about the proteins is their individual structures. If there were sufficient known structures of the protein-protein pairs they could provide direct validation of the interactions. However, the number of such known structures remains small, and certainly nowhere near the number of interacting pairs that have been reported. However, there are relatively large numbers of individual protein structures. This, together with improvements in docking methods makes it possible to begin investigating the likelihood of forming individual three dimensional pairs of structures[8]. Looking at the 3D structure of each protein, especially the binding sites, in an interacting cluster can reveal information that can aid in validating the pair interactions. Some questions that we set out to investigate here are:

1. Whether two proteins prefer to interact

2. If more than two proteins purportedly interact with the same protein, can they interact concurrently by binding two separate regions of the protein, or does one exclude the other because their binding sites substantially overlap?

3. What are the critically important proteins in a protein interaction network?

We chose the yeast protein-protein interaction network. We collect the interaction data from the online database biogrid.com[4]. The number of distinct proteins and interactions in the dataset has increased many folds since Sen *et al*. analyzed the yeast protein network from BioGrid. The current dataset (version 2.0.55) has over five thousand proteins and more than 145,000 interactions.

A.2 Methods

We applied an eigenmode analysis to cluster the protein interaction networks. We form the Kirchhoff matrix[3] M, the interaction matrix, with elements as: for nondiagonal elements of the matrix, M: $M_{ij} = 1$ if i interacts with j and 0 otherwise. For the diagonal elements $M_{ii} = -\sum_{i \neq j} M_{ij}$. Then, we perform eigenmode analysis of this matrix M. By definition, the diagonal elements of the connectivity matrix are the sums of the nondiagonal elements of the given column (or row) taken with the negative sign. This automatically leads to a singular connectivity matrix (i.e. the determinant of the matrix is zero) that must be analyzed with Singular Value Decomposition[2].

A.2.1. Singular Value Decomposition (SVD) computations

We calculate all eigenvalues and eigenvectors of the connectivity matrix by applying the SVD subroutine available in the LAPACK library[9].

If A is any matrix of size $m \times n$ (with m>=n), then A can be written as a product of three matrices:

$$A = U\Lambda V^T \qquad (1)$$

where Λ is the square matrix of size $n \times n$ containing nonnegative values $\lambda_1, \lambda_2, \ldots \lambda_n$ along the diagonal and zeros off diagonal, and U and V are two matrices of sizes $m \times n$ and $n \times n$, respectively, having orthogonal columns, i.e.

$\sum_{i=1}^{m} U_{ik}U_{in} = \delta_{kn}$ and $\sum_{i=1}^{n} V_{ik}V_{in} = \delta_{kn}$ (2)

The Kirchhoff matrix M can be written as

$M = U\Lambda V^T$ (3)

where Λ is the diagonal matrix containing eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_n$ of M and U is the matrix formed from eigenvectors of M. Thus, the elements M_{ij} of the contact matrix M can be expressed as

$M_{ij} = \sum_{k=1}^{n} \lambda_k u_{ki} u_{kj}$ (4)

where u_{ki} denotes the i^{th} component of the eigenvector corresponding to the k^{th} eigenvalue. Equation 4 is the eigenvalue expansion of the contact matrix. From Eq. 4, it follows:

$M_{ii} = \sum_{k=1}^{n} \lambda_k u_{ki}^2$ (5)

The eigenvalues corresponding to the largest absolute values of λ make the largest contributions and smaller eigenvalues contribute successively less [2].

A.2.2. Cluster formation

For each eigenvalue there is a corresponding eigenvector. The significant components of an eigenvector comprise a cluster where each component corresponds to one protein. The components with an absolute value greater than 0.05 are assumed to be significant[2]. The clusters for larger eigenvalues are thus the interesting ones.

A.2.3. Interaction validation

After the clusters are selected, we choose a specific cluster (cluster 14 of network-2.0.41). Then we attempt to test the interactions in this cluster. The steps of this process are shown in the flowchart in Fig. 1. In part (a) an interacting partner protein structure is either retrieved from the protein data bank[10] (www.rcsb.org), or if there is no structure of the protein, then we predict

the structure by comparative modeling. Fig. 1(b) shows that once we have both structures of a putative interacting pair, we then use docking to predict the structure of the interaction complex.

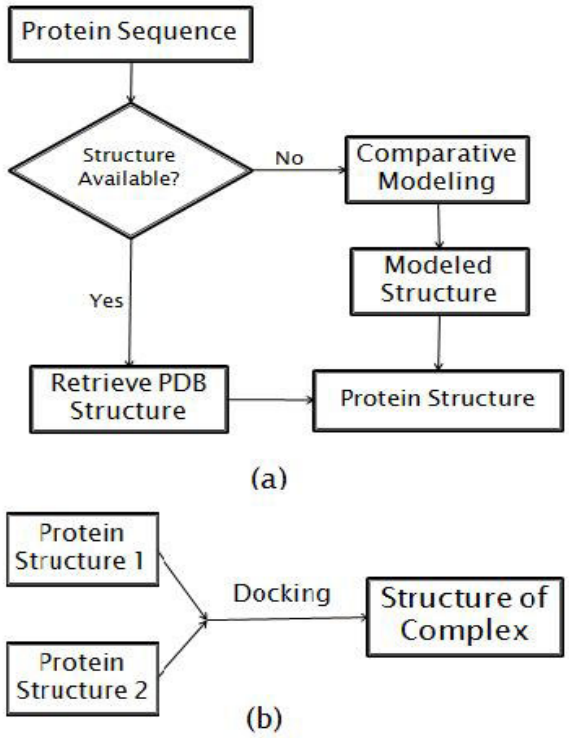


Figure 1. Method for structural validation of a protein interaction pair. (a) Flowchart for obtaining each protein structure (b) Flowchart for docking two proteins to form a docked complex.

A.2.4. Comparative modeling

For structural validation of an interaction, we require the protein structures of both interacting proteins. If the structure of a protein is not available in the protein data bank, we use comparative modeling techniques [11;12]. To predict the structure of the protein, we have relied upon Zhang's I-TASSER server[12-14] (http://zhang.bioinformatics.ku.edu/I-TASSER), which gave the best protein models at the Critical Assessment of Structure Prediction (CASP), a community-wide experiment designed to obtain an objective assessment of the state-of-the-art of the structure prediction field[15-17]. The I-TASSER algorithm consists of three consecutive

steps: threading, fragment assembly, and iteration. During threading, I-TASSER generates the template alignments by a simple sequence Profile-Profile Alignment approach constrained with the secondary structure matches. Fragment assembly is performed on the basis of threaded alignments and the target sequences are divided into aligned and unaligned regions. The fragments in the aligned regions are used directly from the template structures and the unaligned regions are modeled with *ab initio* simulations. Clusters of decoys are generated with the use of a knowledge-based force field. The cluster centroids are generated by averaging the coordinates of all clustered decoys and ranked based on the structure density. In the iteration phase, the steric clashes of the cluster centroids are removed and the topology is refined. The conformations with the lowest energy are selected.

The I-TASSER server returns the best five models with a c-score attached for each model. Also, it returns the top ten templates used in the threading. The c-score is a confidence score that I-TASSER uses to estimate the quality of the predicted model. The calculation of c-score is based on the significance of the threading template alignments and the convergence parameters of the structure assembly simulations. When selecting one of these models, we try to select the model that comes from the largest cluster and that has the highest c-score. C-score is in the range [-5,2], where a higher c-score value signifies a better model[14].

A.2.5. Docking

After we have both structures in an interacting pair we use docking to predict the protein complex formed in a protein-protein interaction. We use the Cluspro server[18-23] for docking the interacting proteins to predict the protein complex. Cluspro is the first fully automated web-based program for docking proteins and was one of the top performers at CAPRI (Critical Assessment of Predicted Interactions), the first community-wide experiment devoted to protein

docking[24]. The Cluspro server is based on a Fast Fourier Transform correlation approach, which makes it feasible to generate and evaluate billions of docked conformations by simple scoring functions. It is an implementation of a multistage protocol: rigid body docking, an energy based filtering, ranking the retained structures based on clustering properties, and finally, the refinement of a limited number of structures by energy minimization. The server (http://cluspro.bu.edu/) returns the top models based on energy and cluster size. We select one of the returned models after considering the energy and the size of the cluster – preferring lower energies and larger cluster sizes. As the Cluspro server implements rigid body docking, when a partner protein in a complex is structurally flexible Cluspro is not so suitable to predict that complex.

A.3 Preliminary Results

We perform the eigen-analysis on the yeast network version 2.0.40 (5,226 proteins and 114,754 interactions) and 2.0.41 (5,425 proteins and 121,664 interactions) and find that the number of zero eigenvalues are 6 and 3, respectively, which are very small compared to those from the yeast network Sen *et al.* previously used[2] (4,906 proteins, 19,037 interactions, and number of zero eigenvalues 46). This decrease in the number of zero eigenvalues is an indication of the completeness of the yeast network. Indeed, in the later versions of the network the number of proteins and interactions has not increased as drastically as previously.

Our analyses of different clusters of the different network versions also support the hypothesis that this type of spectral clustering yields functionally coherent clusters. This is consistent with the findings from some other previous works [2;25].

We have selected three representative clusters (10, 14, and 15) from network-2.0.41 because of their moderate size. The proteins and their interactions in those clusters are shown in Fig. 2.

We note that the number of neighbors for each protein in each of these clusters falls within a relatively small range. Those ranges are 278 – 288 for the proteins in cluster 10; 261 – 286 for the proteins in cluster 14; and 265 – 286 for the proteins in cluster 15.

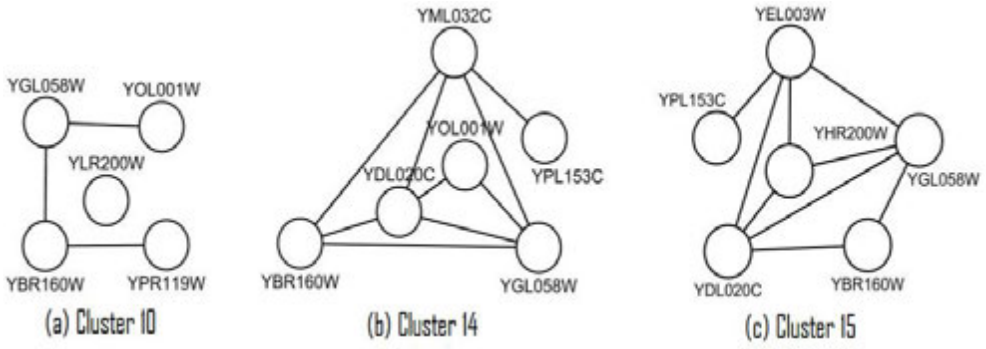


Figure 2. Examples of three clusters and their interactions in clusters 10, 14, and 15, with the nodes being the proteins and their names are given.

We search the gene ontology database[26] for the functions of the proteins in each cluster and find that the proteins in each cluster have related functions for most of the time. This is consistent with the previous findings [2;25]. We also attempt to determine the statistical confidence regarding the functional coherency of the clusters. We used FunSpec[27], a web based cluster interpreter for yeast, to measure the functional coherency of the clusters. The results of our analyses for three of the clusters that we discussed above are shown in Table 1. FunSpec assesses the degree of functional enrichment for a given cluster by the hypergeometric probability distribution[28]. For each cluster, the probability (p-value) of observing such an overlap by chance is calculated as:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

where, G = the size of the genome; C = the number of genes in the genome having that attribute; n = the size of the query cluster; k = the number of gene in the cluster known to have that attribute[28].

Most of the p-values in Table 1 are quite small ($< 10^{-3}$) for the three clusters we are reporting here. These small p-values signify the relatively strong functional coherency of these clusters. How small must a p-value be in order for a cluster to be functionally coherent? FunSpec uses 0.01 as a cut off, which is arbitrary. For each of the clusters, we obtain p-values that are much smaller than 0.01, indicating the highly probable functional coherency of the clusters.

TABLE 1. MIPS functional classification and GO assignments of biological processes and molecular functions for clusters 10, 14, and 15

Cluster #	# proteins	GO molecular function	GO biological process	MIPS functional classification
10	5	Cyclin-dependent protein kinase regulator activity (5×10^{-5}) Tubulin binding(4×10^{-3})	Regulation of cyclin-dependent protein kinase activity (6×10^{-5}) Negative regulation of phosphate metabolic process (9×10^{-4})	Enzymatic activity regulation / enzyme Regulator (5×10^{-4}) Regulation of phosphate metabolism(9×10^{-3})
14	6	Recombinase activity (2×10^{-3}) DNA strand annealing activity (3×10^{-3})	Postreplication repair (1×10^{-4}) regulation of cell cycle (5×10^{-4}) Response to DNA damage stimulus (7×10^{-4})	DNA repair (3×10^{-4}) G2/M transition of mitotic cell cycle (7×10^{-4})
15	6	Protein serine/threonine/tyrosine kinase activity (5×10^{-3})	Regulation of cell cycle (6×10^{-4}) Negative regulation of meiotic cell cycle (10×10^{-4})	Proteasomal degradation (ubiquitin/proteasomal pathway) (2×10^{-4})

One of our goals in this paper is to test the validity of a reported interaction by using structural information about the interacting proteins in a cluster. Our idea is simple: first, find

the structures of the two interacting proteins from the protein data bank[10]. If the experimental structure is not available in the protein data bank for any of the proteins, we predict its structure by comparative modeling. For comparative modeling, we used both CABS modeling[11] and I-TASSER[12-14]. However, the results shown here come only from using I-TASSER. Once, we have both structures, we use docking to predict the docked complex. We can repeat this method to verify individual interaction in a cluster.

Here, we show an example of this approach. We find the homologs for the six proteins in cluster 14 shown in Fig. 2. For the three proteins – YOL001W, YPL153C, and YGL058W – we retrieve the PDB structures having 100% identity as 2PK9 chain B, 1QU5 chain A, and 1AYZ chain A, respectively. For the other three proteins – YBR160W, YML032C, and YDL020C – the PDB homologs are 3EZR chain A (62% identity), 1KN0 chain A (53% identity), and 1A1 chain I (43% identity), respectively. For the latter three proteins, we predict their structures using the I-TASSER server [12-14]. I-TASSER reports the top five predictions for each submitted protein sequence, according to the c-score and the cluster size. We select the model that has the highest c-score out of the five returned models for each target sequence. I-TASSER also returns the top ten templates that it uses for threading. We report the template that has the best sequence identity for the target protein sequence. For each unknown structure, Fig. 4 shows the top prediction, the closest template, and the structural superposition of the predicted structure and the template. The c-scores for the models of YBR160W, YDL020C, and YML032C are 0.65, 0.41, and -0.54, respectively.

We also compute the surface areas for each of the models and the reported template by using NACCESS which is an implementation of the methods described by Lee and Richards[29] and Hubbard, Campbell and Thornton[30]. The surface areas for the model for YBR160W and its

template (PDB id 2PK9A) are 15,727Å^2 and 15,074Å^2, respectively which are close. Also, the surface areas of the model of YDL020C and its template (PDB id 1Z1NX) are 33,655Å^2 and 33,482Å^2, respectively. However, the surface areas for the model for YML032C and its template (PDB id 1W0RA) are 40,965Å^2 and 29,510Å^2, rather different, but this structure does not resemble a usual globular protein. The similarity in these surface areas can serve as a crude indication of the quality of the model returned from the server.

In cluster 14, there are nine interactions. Four interactions involve YML032 whose model returned from the I-TASSER server is not a globular protein. This model is a very extended open structure. As a result, it would appear to have significant structural flexibility and thus not be fully suitable for rigid body docking using Cluspro. We have performed docking for the other five interactions. Results of docking for these five interactions are shown in Fig. 3. For each interaction, the figure shows the surface views of the docked complexes. It is evident from Fig. 3(a), (b), and (c) that protein YDL020C has at least two binding sites. YOL001W and YBR160W both bind to YDL020C at overlapping sites but YGL058W binds with YDL020C at a completely different binding site. Thus, only interactions YDL020C:YBR160W and YDL020C:YGL058W or YDL020C:YOL001W and YDL020C:YGL058W could occur simultaneously.

To measure how strongly these docked complexes are formed, we have calculated the buried surface area for each docked complex. Table 2 shows the buried surface area of each of the docked complexes.

Table 2. Buried surface of the docked complexes+	
Interacting complex	Buried surface (Å^2)
YDL020C : YBR160W	4,295
YDL020C : YOL001W	2,162
YDL020C : YGL058W	4,517
YBR160W : YGL058W	5,603
YOL001W : YGL058W	3,408

+ the order of the complexes in the table is the same as in Fig. 3(a –e).

Complex YBR160W:YGL058W has the largest buried surface area (5,603Å^2) and YDL020C:YOL001W has the smallest (2,162 Å^2). If we consider the buried surface area as a measure of the strength of an interaction between two proteins, the first complex is expected to be more stable than the latter.

A.4 Discussion and Conclusion

This has taken the approach of testing putative protein interactions through their molecular structures. Since not all complexes are available in the Protein Data Bank, nor are they all likely to ever be available, we have relied upon comparative modeling and docking methods. Their recent increased reliability is some justification for these approaches. It has the advantage that it can also identify interactions that could occur together or ones that are mutually exclusive. In addition indirect interactions through another intermediate protein can be identified. However, because of the lengthy computational times and the required human judgment to select models from the results of the prediction programs for comparative modeling and docking, this process cannot yet be fully automated. Nonetheless many such cases can be investigated, and it appears that the results can provide important new information.

This work requires human supervision in each of the computational steps. Although this kind of supervision cannot be completely eliminated, we can automate some of the steps. Model YML032C and its template 1W0RA are highly flexible with probably some disordered regions. New methods that allow combining docking with folding of the disordered parts of a protein structure have been recently proposed [31-34]. We will investigate docking of this category of proteins with its interacting partners in the future work. Moreover, our selected clusters did not contain any interaction where the predicted complex was found in the PDB database which could be because of the fewness of the 3D structures for complexes in the PDB database. By automating some of the steps, we will look into more clusters to see whether some predicted structures are found in the PDB database for experimental validation.

Acknowledgment

We acknowledge the assistance of Taner Z. Sen (Iowa State University) and Michael Zimmermann (L.H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University). We also thank S. Hubbard for generously providing his NACCESS program, which is an implementation of the methods described by Lee and Richards[29] and Hubbard, Campbell and Thornton[30] to compute the buried surface areas of protein complexes.

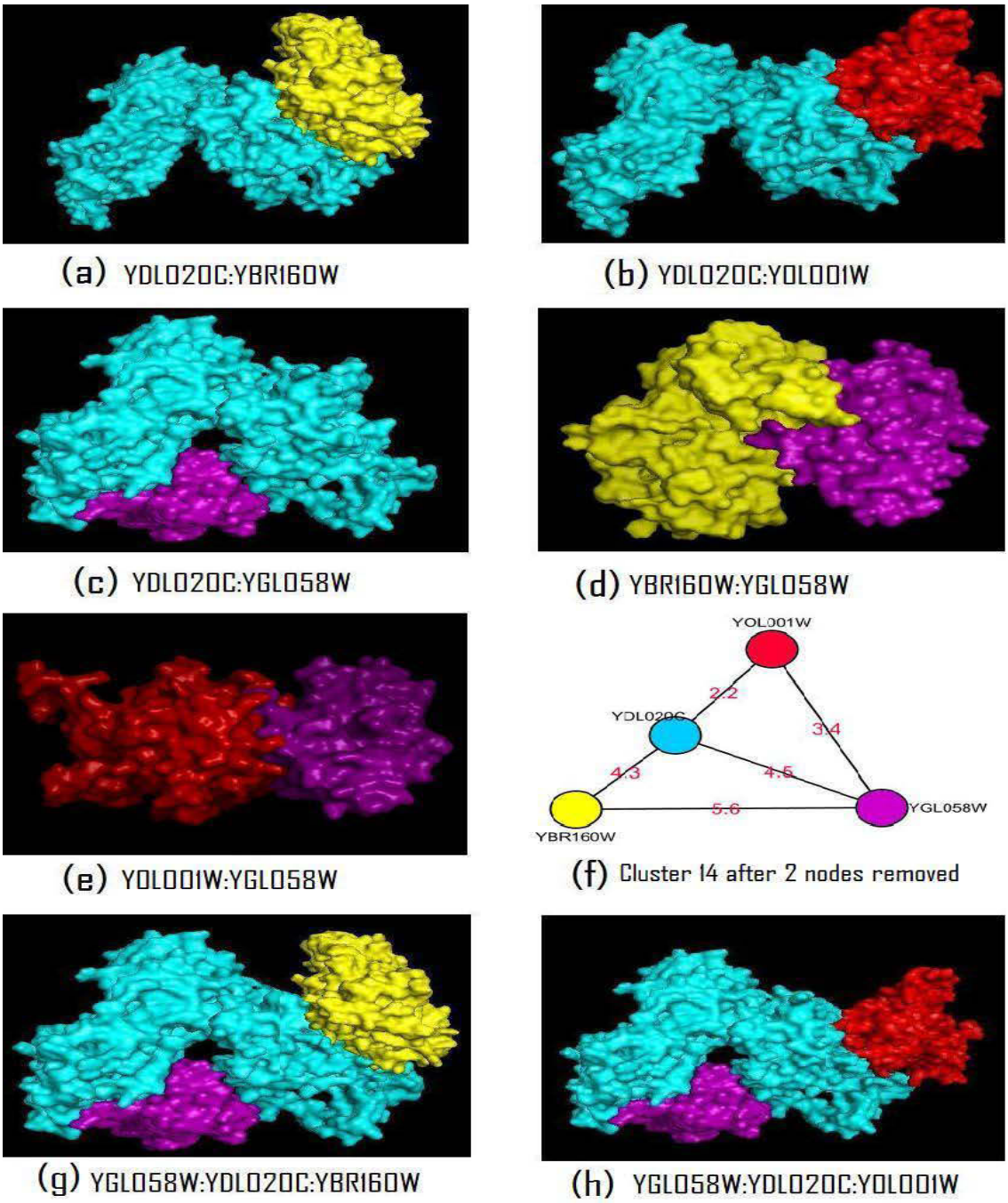


Figure 3. Buried surface areas: (a) 4,295 Å^2 (b) 2,162 Å^2 (c) 4,517 Å^2 (d) 5,603 Å^2 (e) 3,408 Å^2 (g) 8,812 Å^2 (hypothetical complex if interactions a and c occur simultaneously) (h) 6,680 Å^2 (hypothetical complex if interactions b and c occur simultaneously) (f) the proteins in cluster 14 (after YML032C and YPL153C and corresponding edges were removed) – a number on an edge show the buried surface area (in kilo angstroms) of the complex formed by interaction shown by that edge.

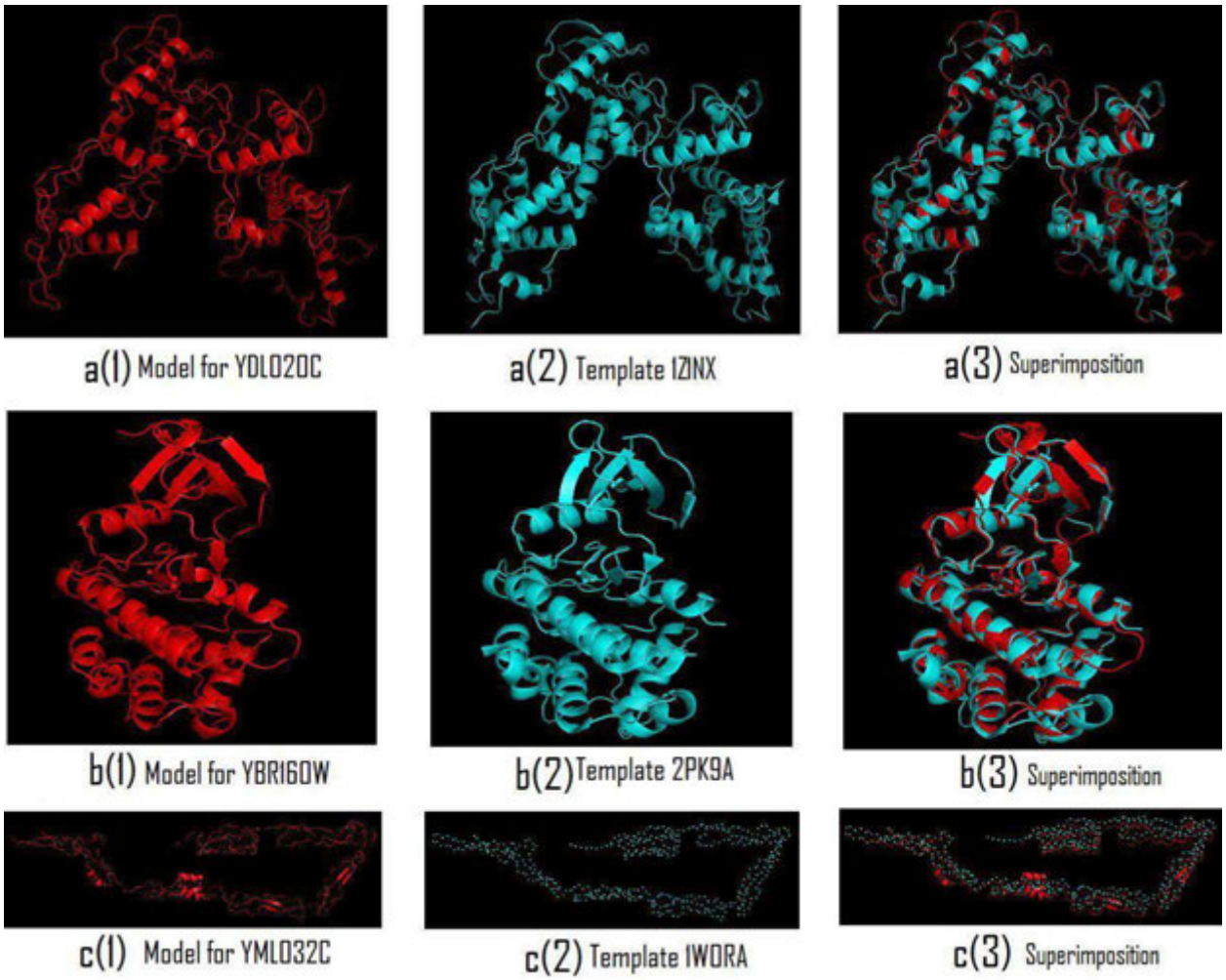


Figure 4. Comparative modeling for three unknown proteins in cluster 14 shown in Fig. 2. a(1) Model for YDL020C a(2) One of the templates used by I-TASSER a(3) Superposition of the model and the template (RMSD = 0.410) b(1) Model for YBR160W b(2) One of the templates used by I-TASSER b(3) Superimposition of the model and the template (RMSD = 0.77) c(1) YML032C c(2) Template used by I-TASSER c(3) Superimposition of the model and the template (RMSD = 0). The difference in buried surface area for the model in a(1) and template in a(2) is 654 Å^2 and that is in b(1) and b(2) 173 Å^2.

Bibliography

[1] K.H.Young, "Yeast Two-Hybrid: So many interactions, (in) So Little Time ....,", 58 ed 2009, pp. 302-311.

[2] T. Z. Sen, A. Kloczkowski, and R. L. Jernigan, "Functional clustering of yeast proteins from the protein-protein interaction network," *BMC. Bioinformatics.*, vol. 7, p. 355, 2006.

[3] S. M. Patra and S. Vishveshwara, "Backbone cluster identification in proteins by a graph theoretical method," *Biophys. Chem.*, vol. 84, no. 1, pp. 13-25, Feb.2000.

[4] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Res.*, vol. 34, no. Database issue, p. D535-D539, Jan.2006.

[5] M. W. Roberts, C. M. Ongkudon, G. M. Forde, and M. K. Danquah, "Versatility of polymethacrylate monoliths for chromatographic purification of biomolecules," *J. Sep. Sci.*, July2009.

[6] P. Aloy and R. B. Russell, "Ten thousand interactions for the molecular biologist," *Nat. Biotechnol.*, vol. 22, no. 10, pp. 1317-1321, Oct.2004.

[7] M. C. von, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399-403, May2002.

[8] P. Aloy and R. B. Russell, "Understanding and Predicting Protein Assemblies With 3D Structures," *Comp Funct. Genomics*, vol. 4, no. 4, pp. 410-415, 2003.

[9] Anderson E, Demmel J, Bai Z, Bischof C, Blackford S, Dongarra J, Du Croz, Greenbaum A, Hammarling S, McKenney A, and Sorensen D, "LAPACK Users' Guide 3rd Edition," 1999.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, Jan.2000.

[11] A. Kolinski, "Protein modeling and structure prediction with a reduced representation," *Acta Biochim. Pol.*, vol. 51, no. 2, pp. 349-371, 2004.

[12] Y. Zhang, "Template-based modeling and free modeling by I-TASSER in CASP7," *Proteins*, vol. 69 Suppl 8, pp. 108-117, 2007.

[13] S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative TASSER simulations," *BMC. Biol.*, vol. 5, p. 17, 2007.

[14] Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC. Bioinformatics.*, vol. 9, p. 40, 2008.

[15] J. N. Battey, J. Kopp, L. Bordoli, R. J. Read, N. D. Clarke, and T. Schwede, "Automated server predictions in CASP7," *Proteins*, vol. 69 Suppl 8, pp. 68-82, 2007.

[16] D. Cozzetto, A. Kryshtafovych, M. Ceriani, and A. Tramontano, "Assessment of predictions in the model quality assessment category," *Proteins*, vol. 69 Suppl 8, pp. 175-183, 2007.

[17] J. Kopp, L. Bordoli, J. N. Battey, F. Kiefer, and T. Schwede, "Assessment of CASP7 predictions for template-based modeling targets," *Proteins*, vol. 69 Suppl 8, pp. 38-56, 2007.

[18] S. R. Comeau, D. Kozakov, R. Brenke, Y. Shen, D. Beglov, and S. Vajda, "ClusPro: performance in CAPRI rounds 6-11 and the new server," *Proteins*, vol. 69, no. 4, pp. 781-785, Dec.2007.

[19] Y. Shen, R. Brenke, D. Kozakov, S. R. Comeau, D. Beglov, and S. Vajda, "Docking with PIPER and refinement with SDU in rounds 6-11 of CAPRI," *Proteins*, vol. 69, no. 4, pp. 734-742, Dec.2007.

[20] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda, "PIPER: an FFT-based protein docking program with pairwise potentials," *Proteins*, vol. 65, no. 2, pp. 392-406, Nov.2006.

[21] S. R. Comeau, S. Vajda, and C. J. Camacho, "Performance of the first protein docking server ClusPro in CAPRI rounds 3-5," *Proteins*, vol. 60, no. 2, pp. 239-244, Aug.2005.

[22] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: a fully automated algorithm for protein-protein docking," *Nucleic Acids Res.*, vol. 32, no. Web Server issue, p. W96-W99, July2004.

[23] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics.*, vol. 20, no. 1, pp. 45-50, Jan.2004.

[24] S. Vajda and D. Kozakov, "Convergence and combination of methods in protein-protein docking," *Curr. Opin. Struct. Biol.*, vol. 19, no. 2, pp. 164-170, Apr.2009.

[25] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Res.*, vol. 31, no. 9, pp. 2443-2450, May2003.

[26] "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) ," 2009.

[27] M. D. Robinson, J. Grigull, N. Mohammad, and T. R. Hughes, "FunSpec: a web-based cluster interpreter for yeast," *BMC. Bioinformatics.*, vol. 3, p. 35, Nov.2002.

[28] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat. Genet.*, vol. 22, no. 3, pp. 281-285, July1999.

[29] B. Lee and F. M. Richards, "The interpretation of protein structures: estimation of static accessibility," *J. Mol. Biol.*, vol. 55, no. 3, pp. 379-400, Feb.1971.

[30] S. J. Hubbard, S. F. Campbell, and J. M. Thornton, "Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors," *J. Mol. Biol.*, vol. 220, no. 2, pp. 507-530, July1991.

[31] I. Coluzza and D. Frenkel, "Monte Carlo study of substrate-induced folding and refolding of lattice proteins," *Biophys. J.*, vol. 92, no. 4, pp. 1150-1156, Feb.2007.

[32] A. G. Turjanski, J. S. Gutkind, R. B. Best, and G. Hummer, "Binding-induced folding of a natively unstructured transcription factor," *PLoS. Comput. Biol.*, vol. 4, no. 4, p. e1000060, Apr.2008.

[33] G. M. Verkhivker, D. Bouzida, D. K. Gehlhaar, P. A. Rejto, S. T. Freer, and P. W. Rose, "Simulating disorder-order transitions in molecular recognition of unstructured proteins: where folding meets binding," *Proc. Natl. Acad. Sci. U. S. A*, vol. 100, no. 9, pp. 5148-5153, Apr.2003.

[34] P. E. Wright and H. J. Dyson, "Linking folding and binding," *Curr. Opin. Struct. Biol.*, vol. 19, no. 1, pp. 31-38, Feb.2009.

APPENDIX B. IMMUNOLOGICAL IMPLICATION OF STRUCTURAL ANALYSIS OF PORCINE IL1β PROTEINS EXPRESSED IN MACROPHAGES AND EMBRYOS

Modified from a published manuscript in the peer reviewed conference proceedings, ACM International Conference on Bioinformatics and Computational Biology, August 2-4, 2010. Niagara Falls, New York, USA

Ataur R Katebi, Pawel Gniewek, Michael Zimmermann, Saras Saraswathi, Zhenming Gong, Christopher K. Tuggle, Andrzej Kloczkowski, Robert L. Jernigan

Abstract

IL1β is an important vertebrate animal protein. It is a member of the cytokine protein family and is involved in generating an inflammatory response to some infections. Researchers have found that there are two porcine IL1β proteins expressed – one in embryos and the other in macrophage and endometrial tissues. However, these two proteins have about 86% sequence identity. In this paper, we attempt to describe how these two proteins might be structurally and functionally different. We find interesting aspects of these two structures that differ: 1) A predicted binding site appears to have different side chain arrangements that might lead to different binding efficiencies for the same protein or even to different partners. 2) The Caspase 1 cleavage site in the precursor proteins differs in a way that has previously been experimentally determined to be important and to reduce the cleavage activity by one order of magnitude for the embryonic IL1β, conferring a significant advantage to the protein (embryonic IL1β).

B.1 Introduction

Living organisms possess immune systems ranging from simple selective membranes to biological entities that can detect and protect the organism against disease-causing pathogens or

rogue tumor cells. Vertebrates such as humans and pigs have evolved a complicated system of defense mechanism against pathogens. Leukocytes, also known as white blood cells, are one such part of the immune system. Macrophages are a type of such leukocytes that engulfs foreign bodies (phagocytosis) and isolates them within cell membranes. Cells in the immune system, including other types of leukocytes, secrete substances called cytokines. These cytokines signal neighboring cells, eliciting a response to the invading pathogen. This signaling results in the production of certain proteins or peptides which help fight the invading pathogens. Interleukins (IL), representing a major part of the immune system, belong to one group of cytokines. Interleukins form the primary communication channels of the immune system. The interleukins are synthesized by leukocytes such as macrophages and endothelial cells such as the inner lining of the mammalian uterus (endometrium).

There are many types of interleukins with varying functions and effects on the immune system. We are interested in IL1β, which is an inflammatory cytokine activated by microphages as part of the immune response to infection. It increases cell-surface adhesion by helping to synthesize cell-surface adhesion molecules and helps recruit leukocytes to the infected site. In the blood stream it can cause fever and helps synthesize proteins, that can activate transcription factors to stimulate gene expression.

IL1β is an important vertebrate animal protein. It is a member of the IL-1 cytokine protein family which is produced by activated macrophages as a proprotein. The IL1β proprotein is proteolytically cleaved to its active form by Caspase 1. In humans, this protein is related to many diseases such as major depressive disorder[1], osteoporosis in post-menopausal women [2], lung cancer in a Japanese population [2], increased bleeding after cardiac surgery [3], chronic and aggressive periodontitis [4], chronic inflammatory conditions of the brain [5], etc. This protein

also takes part in a variety of cellular activities such as cell proliferation, differentiation, and apoptosis.

The porcine genome has just been sequenced. We followed up on some prior QPCR (quantitative real time polymerase chain reaction) and sequencing work that indicated there were two different IL1β RNAs expressed. Two genomic copies appear to be tandem duplicates of IL1β. The RNA sequence data indicates that one is expressed in macrophages and endometrium tissues and the other is expressed in the developing embryo at implantation. These two predicted protein sequences are 86% identical. However, in the embryonic case, there is a proline inserted just 2 amino acid away from the predicted Caspase 1 cleavage site that activates the protein.

Here we seek answers to the following questions:

- How are these two proteins structurally different from one another?
- Do these two proteins have any apparent differences in functions?

The tyrosine kinase family that IL1β belongs to, has two sub-classes: receptor and non-receptor. Receptor class proteins play pivotal roles in diverse cellular activities including growth, differentiation, metabolism, adhesion, motility, and death. So it is possible that the activity of the kinases in embryos and macrophages might be different. We might suppose that this activity could be higher in embryos than in macrophages. In this paper we present our findings to support the idea that despite small differences between the sequences of these two proteins, mIL1β (from macrophages) and ayIL1β (from embryos), there may be some significant differences in the activity of these two proteins.

B.2 Methods

B.2.1 Comparative Modeling

To predict an interaction complex or predict a new interaction, we require the protein structures of both interacting proteins. If the structure of a protein is not available in the PDB, we use comparative modeling approaches [6;7]. For structure prediction of the proteins, we have relied upon Zhang's I-TASSER server[7-9] (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) to predict the structures of mIL1β and ayIL1β as well as their corresponding receptors; IL1R1, IL1R2, and IL1RN. This algorithm gave the best protein models at the most recent Critical Assessment of Structure Prediction (CASP 7 and CASP 8), a community-wide, worldwide experiment designed to obtain an objective assessment of the state-of-the-art methods in structure prediction[10-12]. The I-TASSER algorithm consists of three consecutive steps: threading, fragment assembly, and iteration. During threading, I-TASSER generates template alignments by a simple sequence Profile-Profile Alignment approach constrained with the secondary structure matches. Fragment assembly is performed on the basis of threaded alignments and the target sequences are divided into aligned and unaligned regions. The fragments in the aligned regions are used directly from the template structures and the unaligned regions are modeled with *ab initio* simulations. Clusters of decoys are generated with the use of a knowledge-based force field. The cluster centroids are generated by averaging the coordinates of all clustered decoys and ranked based on structure density. In the iteration phase, the steric clashes of the cluster centroids are removed and the topology is refined. The conformations with the lowest energy are selected.

The I-TASSER server returns the best five models with a C-score attached for each model and the top ten templates used. The C-score is a confidence score that I-TASSER uses to

estimate the quality of the predicted model. The calculation of C-score is based on the significance of the threading template alignments and the convergence parameters of the structure assembly simulations. When selecting one of these models, we select the model that comes from the largest cluster and has the best C-score. Reasonable structures are usually found to have a C-score in the range [-5,2], where a higher C-score value signifies a better (more confident) model[9].

B.2.2 Docking

After we have structural models for each protein, we use docking to predict the protein complex formed in the cytokine-receptor pairs. We use the Cluspro server [13-18] for docking the interacting proteins. Cluspro is the first fully automated web-based program for docking proteins and was one of the top performers at CAPRI (Critical Assessment of Predicted Interactions) rounds 1-12, the community-wide experiment devoted to protein docking[19]. The Cluspro server is based on a Fast Fourier Transform correlation approach, which makes it feasible to generate and evaluate billions of docked conformations by simple scoring functions. It is an implementation of a multistage protocol: rigid body docking, an energy based filtering, ranking the retained structures based on clustering properties, and finally, the refinement of a limited number of structures by energy minimization. The server (http://cluspro.bu.edu/) returns the top models based on energy and cluster size. We select one of the returned models after considering the energies and the sizes of the clusters – preferring lower energies and larger cluster sizes. As the Cluspro server implements rigid body docking, then in cases when a partner protein in a complex is structurally flexible, Cluspro would not be able to account for this flexibility.

B.3 Preliminary Results

B.3.1 Structure Similarity

We predict the structures of the cores of the two porcine proteins, mIL1β and ayIL1β, and find that structurally they are closely similar with an all atom RMSD value 0.48 Å as shown in Figure 1. We also predict the precursor structures of these two porcine proteins. The C-scores for the structures of the precursors of mIL1β and ayIL1β are -3.07 and -3.06, respectively. These low C-scores indicate that structure prediction for the leading sequences of the precursor proteins was mostly *ab initio* structure predictions. We separately consider the core part from each of the precursor structures and superimpose them as shown in Figure 2. The all atom RMSD value is 1.61 Å, which is significantly larger than 0.48 Å found when the cores were predicted independently. It is possible that these larger differences in the excised leader part might have some influence on the folding of the core proteins.

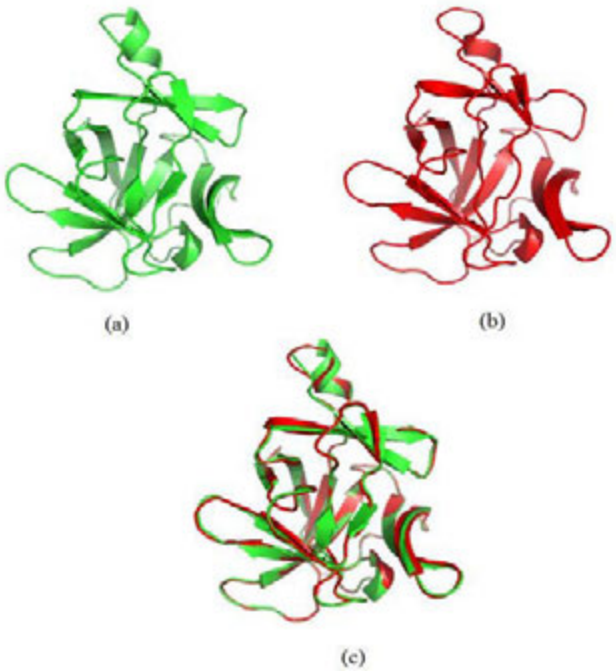


Figure 42. Comparison of the predicted structures of the two porcine IL1β structures. (a) Predicted structure of mIL1B with C-score 1.76 (b) Predicted structure of aylL1 β with C-score 1.65 (c) Super imposition of the predicted structures in (a) and (b); all atom RMSD = 0.48 Å.

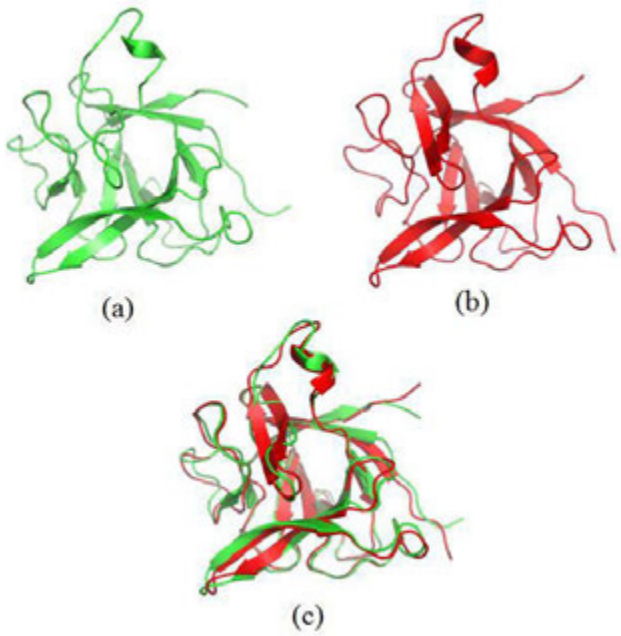


Figure 43. Comparison of the caspase excised parts of the two IL1β structures. (a) Structure of the mature part extracted from predicted precursor structure for mIL1β protein (b) Structure of the mature part extracted from predicted precursor structure for aylL1β protein (c) Super imposition of the structures in (a) and (b); RMSD = 1.61 Å.

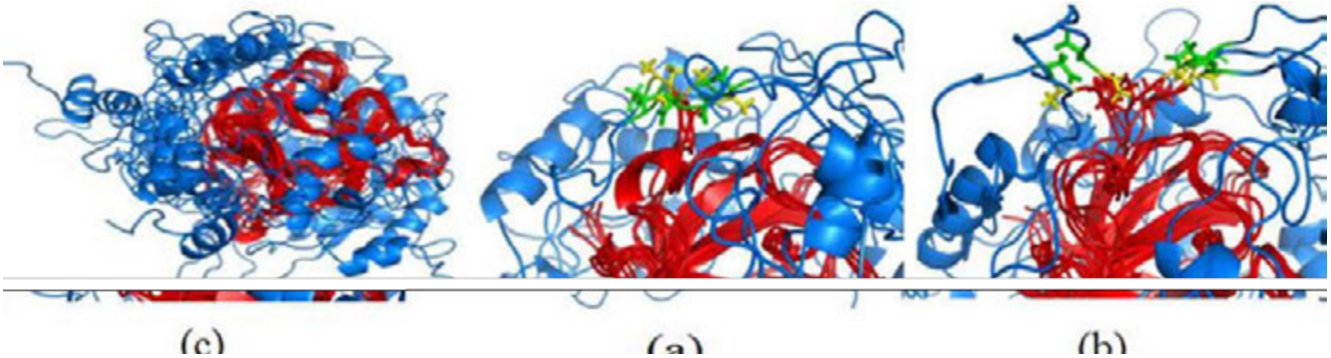


Figure 44. Comparison of predicted porcine IL1β structures with the human IL1β. (a) We superimpose onto 1ITBA the 5 best ITASSER homology models of each of the two porcine IL1β pre-cleavage protein sequences using the CE algorithm. We see that for all of the models the core (red) is predicted to have nearly the same fold while the leading sequence (blue) is placed in many configurations. We show (b) the 5 mIL1β sequence models and (c) the 5 ayIL1β sequence models with the proline at the cleavage site as sticks. There does not appear to be a large effect on the Caspase 1 cleavage site caused by the introduction of the proline.

We have compared the predicted structures for the precursor sequences of mIL1β and ayIL1β with the experimental structure of human IL1β. The PDB (Protein Data Base) structure, 1ITB, is a structure for type-1 interleukin-1 receptor complexed with interleukin-1 beta (IL1β). We have separated the structure (chain A) of human IL1β from this co-structure. When we superimpose the 5 best I-TASSER homology models of each of the two porcine IL1β pre-cleavage protein sequences onto 1ITBA, we see that for all of the models the core is predicted to have nearly the same structure while the structure of the leading sequence is placed in many configurations as shown in Figure 3.

The presence of a disordered part in a protein structure make the dynamics of the protein less constrained. Linker regions are important for diverse purposes, ranging from viral attachment proteins to transcription factors. The disordered regions in a protein structure also facilitates protease digestion[20]. We have predicted the disorder of the two protein sequences using different prediction servers [21;22] and found that most prediction servers give the cleavage site

of these two proteins as disordered as shown in Figure 4 (113 in mIL1β and 114 in ayIL1β). The most disordered regions are in the core of these two proteins are loops connecting strands in beta sheets.

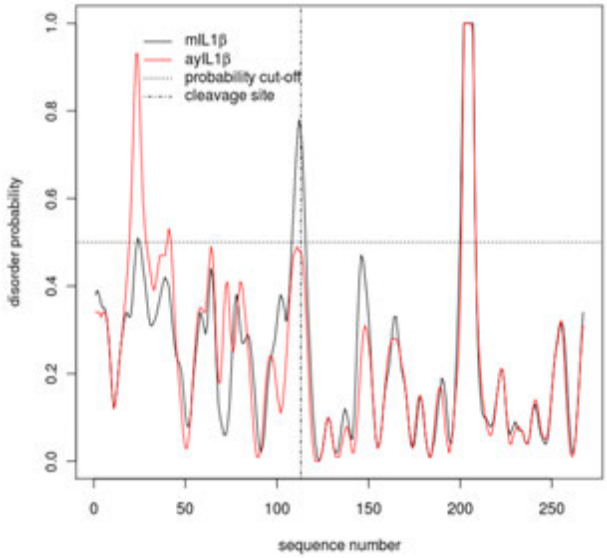


Figure 45. Disorder prediction for (a) mIL1β and (b) ayIL1β proteins.

B.3.2 Function Similarity

We have also some observations about the predictions of the binding sites and functions of these two proteins. First we used sequence based methods for function prediction of the two proteins using the ExPaSy server (http://ca.expasy.org/prosite). This server gives the motif, interleaukin-1, for each of the sequences. For the mIL1β protein sequence the amino acids 112-132 form the motif and for ayIL1B the motif region is 113-133. In addition to this prediction, we have the same functional site predicted using the I-TASSER server. Both proteins are predicted to have two functions with EC-score>1.1. A functional prediction with EC-score>1.1 is considered to be a highly confident prediction. One of these predicted functions is the enzymatic activity in the following phosphorylation reaction:

ATP + a [protein]-L-tyrosine <=> ADP + a [protein]-L-tyrosine phosphate

The other function is the limited hydrolysis of proteins of the neuroexocytosis apparatus but no action is detected on small molecule substrates. The first function is predicted with a higher level of confidence. Therefore, we focus only on the first function, the enzymatic activity.

Based on the prediction of binding sites from I-TASSER for the mIL1β protein we have predicted two possible binding sites with low probability:

Binding site 1 (BS-score=0.36): 18(LEU), 19(VAL), 20(LEU), 21(ALA), 22(GLY), 23(PRO), 29(LEU), 31(LEU), 35(ASP), 36(LEU), 37(LYS), 38(ARG), 39(GLU), 40(VAL), 62(ILE), 65(LYS), 129(GLN)

Binding site 2 (BS-score=0.24): 123(SER), 124(THR), 126(GLN), 133(PHE), 135(GLY), 137(SER), 138(LYS), 139(GLY), 140(ARG), 141(GLN), 142(ASP), 143(ILE), 144(THR)

Notably these differ somewhat between mIL1β and ayIL1β. The predictions for the two binding sites for ayIL1b are as follow:

Binding site 1 (BS-score=0.38): 19(LEU), 20(VAL), 21(LEU), 22(ALA), 23(GLY), 24(PRO), 30(LEU), 32(LEU), 36(ASP), 37(LEU), 38(LYS), 39(ARG), 40(GLU), 41(VAL), 63(ILE), 66(LYS)

Binding site 2 (BS-score=0.55): 124(SER), 125(THR), 126(SER), 127(GLN), 132(PRO), 134(PHE), 138(SER), 142(GLN), 143(ASP), 144(ILE)

Based on the templates used for this prediction, we can conclude that binding site 1 in ayIL1β probably does not play an important role in “tyrosine kinase” function. Therefore we ignore this binding site for further consideration and focus only on the second binding site.

The next step in our analysis is explaining the sources of possible different activity between these two proteins. The Cα RMSD of the structures of the two mature proteins is small (0.39 Å), but if we compare positions of binding site residues of mIL1β and ayIL1β, we see a completely different packing of large residues (see Figure 5). Even a small difference in the sequence composition can have a strong effect on protein interactions because of different positional orientation of binding site residues. The orientation of ayIL1β binding site residues is different from the orientation of binding site residues in mIL1β, and thus packed differently to bind with the ligand. It can be proved that [23]:

$$\text{affinity}_1\text{-affinity}_2 \sim -(\Delta F_1 - \Delta F_2)/RT \qquad (1)$$

where ΔF is the difference in the free energy of binding of two molecules (the mIL1β and the ligand, and the ay IL1β and the ligand).

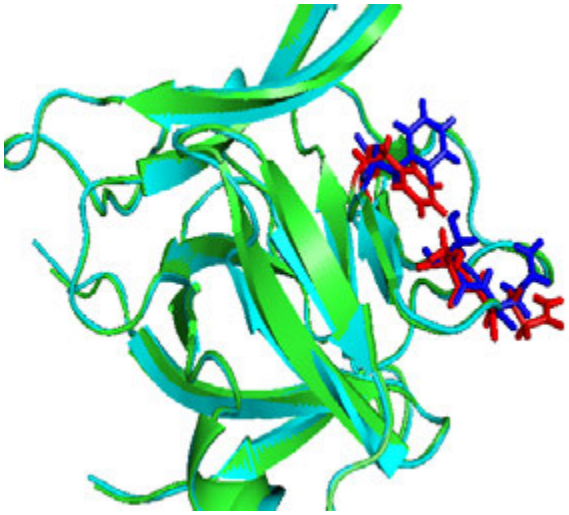


Figure 46: Superposition of the models for the cores of mIL1β and ayIL1β with some of the side chain residues in the predicted 2nd binding site for each. Red: mIL1β, Blue: ayIL1β.

If we were able to calculate the right side of equation 1, we can assess the relative affinity (binding strength) for mIL1β and ayIL1β with the ligand. Binding can occur in both cases. But

the affinity is probably different in these cases, and thus the biological response on expression may be different. This may indicate why ayIL1β is expressed in embryo and mIL1β in macrophage.

We also predicted the post translational modification for the mature parts of these two proteins.

The modification predicted to occur in the 2nd binding site of mature ayIL1β protein but is not predicted for mature mIL1β. The 133rd residue, which is a proline in mature ayIL1β, is predicted to be hydroxylyzed after translation and forms hydroxyproline. We should keep in mind that the score is quite low, 0.30 (http://ams2.bioinfo.pl/).

We can see that in case of the template structure, which was used to predict the 2nd binding site, hydroxylyzed proline comes in closer contact with the nitrogen atom of the ligand as shown in Figure 6.

Because of this hydroxyproline modification (Post translational modification - PTM) in mIL1β and ayIL1β, the positive H atom of hydroxylyzed proline has the opportunity to form additional hydrogen bond with potential ligand and therefore binding such ligand will be stronger in this binding site.

Therefore any post translational modification at the binding site residues activate/deactivate the activity of the proteins.

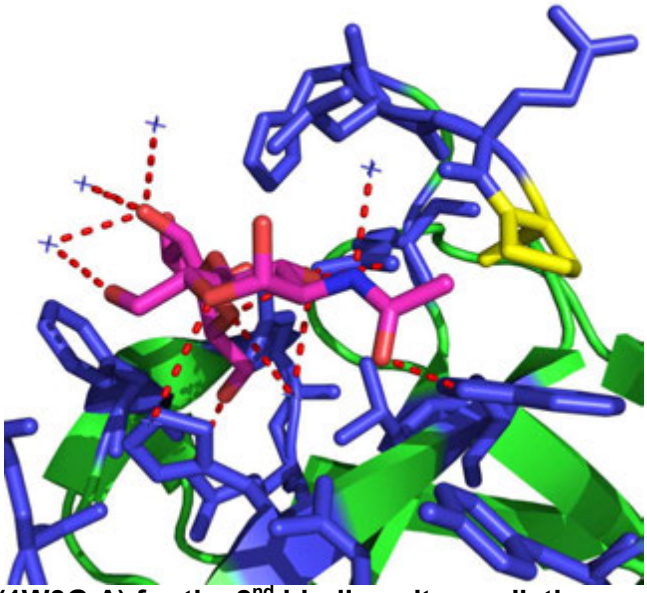


Figure 47. The template (1W3G A) for the 2nd binding site prediction and its ligand (NLC).

B.4 Discussion and Conclusion

We have used I-TASSER to predict the structures of the two porcine IL1β proteins. As these two proteins have high sequence similarity with human IL1β proteins, for verification purposes, we predicted the structure for human IL1β using the I-TASSER server and the RMSD for this predicted structure against the experimental structure of human IL1B is 0.332 Å which is quite low. This says that the structure predictions for the two porcine IL1βs are likely to be reliable.

It is always useful to utilize predicted structures for building new hypotheses. In this case, we are able to postulate two possible hypotheses.

B.4.1 Hypothesis 1

In order to achieve their their biological function, IL1β proteins need to be cleaved [24]. We also found that for these proteins an enzyme that cuts IL1β sequences is ICE (IL1β converting

enzyme) [25]. The two porcine proteins, mIL1β and ayIL1β, have cleavage sites differing slightly in sequence - FVCD*ANVQ and FLCD*ATPV (* indicates cleavage site). It was experimentally found that the right side (ANVQ or ATPV) of a cleavage site does not play an important role, and therefore only left side is the important one [26]. We can see that for these two proteins, the only such difference in the cleavage site is between leucine(L) and valine(V). In mIL1β sequence, this position is occupied by valine (V) and in ayIL1β it is leucine (L). The substrate activity research shows that V in this position is highly promoted, and the activity of mutants with any other amino acid in this position is significantly lower (~one order of magnitude)[27]. It can be explained by a bigger size of the side chain of L in comparison with V, which can lead to worse packing of the binding site of ICE. More efficient cleavage of the precursor for the macrophage, mIL1β, would mean a greater abundance of the mature form.

B.4.2 Hypothesis 2

Embryos and macrophages are two different stages of the same organism. In each case, the set of performed biochemical pathways or their productivity could be different. So in such cases, there is some possibility to control the efficiency of the pathways. One possibility is to modify the activity of the enzymes by changing their binding sites which might inhibit such an enzyme's activity or change its biological activity. Therefore, because the mIL1β and ayIL1β proteins differ in the 2nd predicted binding site and one of them is predicted to have a modified amino acid (133rd Proline in ayIL1β) in the same binding site; this can affect the selective activity between the two, which may be the reason why they are expressed differentially in these two cases.

Other hypotheses are of course possible.

B.5 Acknowledgment

We like to thank Dr. Drena Dobbs for preliminary discussions. Grant sponsor: NIH; Grant numbers: R01GM073095, R01GM072014, and R01GM081680.

Bibliography

[1] **B. T. Baune, U. Dannlowski, K. Domschke, D. G. Janssen, M. A. Jordan, P. Ohrmann, J. Bauer, E. Biros, V. Arolt, H. Kugel, A. G. Baxter, and T. Suslow, "The interleukin 1 beta (IL1B) gene is associated with failure to achieve remission and impaired emotion processing in major depression," *Biol. Psychiatry*, vol. 67, no. 6, pp. 543-549, Mar.2010.**

[2] **B. Czerny, A. Kaminski, M. Kurzawski, D. Kotrych, K. Safranow, V. Dziedziejko, A. Bohatyrewicz, and A. Pawlik, "The association of IL-1beta, IL-2, and IL-6 gene polymorphisms with bone mineral density and osteoporosis in postmenopausal women," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 149, no. 1, pp. 82-85, Mar.2010.**

[3] **I. J. Welsby, M. V. Podgoreanu, B. Phillips-Bute, R. Morris, J. P. Mathew, P. K. Smith, M. F. Newman, D. A. Schwinn, and M. Stafford-Smith, "Association of the 98T ELAM-1 Polymorphism with Increased Bleeding After Cardiac Surgery," *J. Cardiothorac. Vasc. Anesth.*, Jan.2010.**

[4] **A. R. Shete, R. Joseph, N. N. Vijayan, L. Srinivas, and M. Banerjee, "Association of single nucleotide gene polymorphism at interleukin-1beta +3954, -511, and -31 in chronic periodontitis and aggressive periodontitis in Dravidian ethnicity," *J. Periodontol.*, vol. 81, no. 1, pp. 62-69, Jan.2010.**

[5] **B. D. Griffin and P. N. Moynagh, "Persistent interleukin-1beta signaling causes long term activation of NFkappaB in a promoter-specific manner in human glial cells," *J. Biol. Chem.*, vol. 281, no. 15, pp. 10316-10326, Apr.2006.**

[6] **A. Kolinski, "Protein modeling and structure prediction with a reduced representation," *Acta Biochim. Pol.*, vol. 51, no. 2, pp. 349-371, 2004.**

[7] **Y. Zhang, "Template-based modeling and free modeling by I-TASSER in CASP7," *Proteins*, vol. 69 Suppl 8, pp. 108-117, 2007.**

[8] **S. Wu, J. Skolnick, and Y. Zhang, "Ab initio modeling of small proteins by iterative TASSER simulations," *BMC. Biol.*, vol. 5, p. 17, 2007.**

[9] **Y. Zhang, "I-TASSER server for protein 3D structure prediction," *BMC. Bioinformatics.*, vol. 9, p. 40, 2008.**

[10] **J. N. Battey, J. Kopp, L. Bordoli, R. J. Read, N. D. Clarke, and T. Schwede, "Automated server predictions in CASP7," *Proteins*, vol. 69 Suppl 8, pp. 68-82, 2007.**

[11] **D. Cozzetto, A. Kryshtafovych, M. Ceriani, and A. Tramontano, "Assessment of predictions in the model quality assessment category," *Proteins*, vol. 69 Suppl 8, pp. 175-183, 2007.**

[12] **J. Kopp, L. Bordoli, J. N. Battey, F. Kiefer, and T. Schwede, "Assessment of CASP7 predictions for template-based modeling targets," *Proteins*, vol. 69 Suppl 8, pp. 38-56, 2007.**

[13] **S. R. Comeau, D. Kozakov, R. Brenke, Y. Shen, D. Beglov, and S. Vajda, "ClusPro: performance in CAPRI rounds 6-11 and the new server," *Proteins*, vol. 69, no. 4, pp. 781-785, Dec.2007.**

[14] Y. Shen, R. Brenke, D. Kozakov, S. R. Comeau, D. Beglov, and S. Vajda, "Docking with PIPER and refinement with SDU in rounds 6-11 of CAPRI," *Proteins*, vol. 69, no. 4, pp. 734-742, Dec.2007.

[15] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda, "PIPER: an FFT-based protein docking program with pairwise potentials," *Proteins*, vol. 65, no. 2, pp. 392-406, Nov.2006.

[16] S. R. Comeau, S. Vajda, and C. J. Camacho, "Performance of the first protein docking server ClusPro in CAPRI rounds 3-5," *Proteins*, vol. 60, no. 2, pp. 239-244, Aug.2005.

[17] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: a fully automated algorithm for protein-protein docking," *Nucleic Acids Res.*, vol. 32, no. Web Server issue, p. W96-W99, July2004.

[18] S. R. Comeau, D. W. Gatchell, S. Vajda, and C. J. Camacho, "ClusPro: an automated docking and discrimination method for the prediction of protein complexes," *Bioinformatics.*, vol. 20, no. 1, pp. 45-50, Jan.2004.

[19] S. Vajda and D. Kozakov, "Convergence and combination of methods in protein-protein docking," *Curr. Opin. Struct. Biol.*, vol. 19, no. 2, pp. 164-170, Apr.2009.

[20] A. K. Dunker, I. Silman, V. N. Uversky, and J. L. Sussman, "Function and structure of inherently disordered proteins," *Curr. Opin. Struct. Biol.*, vol. 18, no. 6, pp. 756-764, Dec.2008.

[21] S. M. B. P. Cheng J, "Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data,", 11 ed 2005, pp. 213-222.

[22] "Order/Disorder Prediction With Self Organising Maps,". MacCallum B., Ed.

[23] C. F. Karney, J. E. Ferrara, and S. Brunner, "Method for computing protein binding affinity," *J. Comput. Chem.*, vol. 26, no. 3, pp. 243-251, Feb.2005.

[24] N. A. Thornberry and S. M. Molineaux, "Interleukin-1 beta converting enzyme: a novel cysteine protease required for IL-1 beta production and implicated in programmed cell death," *Protein Sci.*, vol. 4, no. 1, pp. 3-12, Jan.1995.

[25] Nancy A.Thornberry and Susan M.Molineaux, "Interleukin-1b converting enzyme: A novel cysteine protease required for IL-1b production and implicated in programmed cell death,", 4 ed 1994.

[26] N. A. Thornberry, H. G. Bull, J. R. Calaycay, K. T. Chapman, A. D. Howard, M. J. Kostura, D. K. Miller, S. M. Molineaux, J. R. Weidner, J. Aunins, and ., "A novel heterodimeric cysteine protease is required for interleukin-1 beta processing in monocytes," *Nature*, vol. 356, no. 6372, pp. 768-774, Apr.1992.

[27] D. K. Miller, J. R. Calaycay, K. T. Chapman, A. D. Howard, M. J. Kostura, S. M. Molineaux, and N. A. Thornberry, "The IL-1 beta converting enzyme as a therapeutic target," *Ann. N. Y. Acad. Sci.*, vol. 696, pp. 133-148, Nov.1993.

APPENDIX C. THE IMPORTANCE OF SLOW MOTIONS FOR PROTEIN FUNCTIONAL LOOPS

This is a published manuscript in the peer reviewed scientific journal, *Physical Biology*

Aris Skliros, Michael T. Zimmermann, Debkanta Chakraborty, Saras Saraswathi, Ataur R. Katebi, Sumudu P. Leelananda, Andrzej Kloczkowski, and Robert L. Jernigan

Abstract

Loops in proteins connect secondary structures such as alpha-helix and beta-sheet, are often on the surface, and may play a critical role in some functions of a protein. The mobility of loops is central for the motional freedom and flexibility requirements of active-site loops and may play a critical role for some functions. The structures and behaviors of loops have not been much studied in the context of the whole structure and its overall motions, and especially how these might be coupled. Here we investigate loop motions by using coarse-grained structures (C^{α} atoms only) to solve for the motions of the system by applying Lagrange equations with elastic network models to learn about which loops move in an independent fashion and which move in coordination with domain motions, faster and slower, respectively. The normal modes of the system are calculated using eigen-decomposition of the stiffness matrix. The contribution of individual modes and groups of modes are investigated for their effects on all residues in each loop by using Fourier analyses. Our results indicate overall that the motions of functional sets of loops behave in similar ways as the whole structure. But, overall only a relatively few loops move in coordination with the dominant slow modes of motion, and that these are often closely related to function.

C.1 Introduction

The importance of understanding loops in proteins

Protein motions are extremely important for their functioning, and it is now well established that domain motions are the dominant motions and that these motions are often relatable to their functions. So immediately there the question arises of whether loops are controlled in their motions by the domain motions and are slow, or whether they move independently and thus more rapidly. The first case, where the loops move together with the large domain motions, corresponds to the protein structure being strongly cooperative in its motions, and these motions will reflect a type of allostery. Distinguishing between these two extremes is the intention of the present study of the behavior of protein surface loops.

One category of loops whose function is clear are those large loops that cover binding sites, and are clearly important since they have to open in order to facilitate binding, and thus these motions are clearly critical for function. But, other loops may have functional roles such as chaperoning the transport of ligands from secondary binding sites towards their primary binding site. Such behaviors might reflect a more deterministic behavior than is usually observed in molecular simulations. New ways of investigating loop behaviors may assist in our understanding of such biased, non-random behavior in biology.

If loops that are functionally important move in a strongly coordinated way with the larger domain motions, i.e., the slowest motions, then it might even be possible to predict which loops are likely to be functional based on computations that identify them as moving more slowly. In addition there is the issue of how the loops move with respect to the domain to which they are attached. If they are fully coordinated in a positive way then they are moving as if the domain

and the loop together were a rigid block. If they are moving in a strongly anti-correlated way, this would require articulated joints between the domain and the loop, but whether correlated or anti-correlated both could be motions effectively under the control of the domain motions.

First we must define loops for the purpose of this investigation. Proteins consist primarily of three types of secondary structure elements: α-helices, β–strands (forming either parallel or antiparallel β-sheets) and loops. Loop regions here are taken to be those conformationally irregular fragments of the chain, that connect between two secondary structure elements and lie upon the surface.

Loops are also quite variable in their lengths and sometimes there are even gaps in the loop regions of some reported protein structures, because of disorder. Godzik and collaborators (1) recently surveyed the PDB to find ordered-disordered pairs; residues of the same protein in two different crystals, where the atomic coordinates were resolved in one, but not in the other. They found that this type of disorder (sometimes relates to post-translational modification) is overrepresented in loop regions (46% of ordered-disordered pairs). While it might be tempting to interpret the missing parts of loops as moving independently, it seems likely that the details of the crystal packing would be likely to confound such a simple interpretation. Completing the structures of these missing regions in loops and learning about the range of loop conformations are important issues not yet a standard computation, that will not be considered here, even though these issues are important for evaluating the importance of the mobilities of all loops. Solving these issues would however result in an improved understanding of the functional roles played by loops.

Diverse approaches have been applied to fill in the missing information in loop regions. Joosten *et al.* (2) have combined structural and electron density information to find likely

conformations of loop regions. Fiser *et al.* (3) presented an improved and automated modeling technique for loop predictions using spatial information and optimization of a pseudo-energy function. Felts *et al.* (4) predicted native conformations using Optimized Potentials for Liquid Simulations, all atom (OPLS-AA) force fields, and Analytical Generalized Born plus Non-Polar (AGBNP) implicit solvent models, in combination with torsion angle conformation based search. By understanding the relationship between the motions of loops and larger domains, it might even be possible to improve crystallographic refinements of loop regions. Importantly, a full comprehension would permit predictions of ensembles of loop conformations, rather than static structures.

Sellers *et al.* (5) predicted loops in inexact environments by examining how loop refinement accuracy is affected by errors in the surrounding elements such as backbone and side-chain positions. They used augmented loop prediction methods that optimize the conformations of the side chains simultaneously. This method helps to recover near-native conformations for many perturbed structures. Olson *et al.* (6) examined *ab initio* methods for predicting protein loops by using multi-scale conformational sampling. They used physical energy functions to score the models. Peng and Yang (7) developed a knowledge-based loop prediction method without the necessity of constructing hierarchically clustered length-dependent loop libraries. This method first predicts the local structure of the loop and then structurally aligns it against all possible motif templates. Zhu *et al.* (8) have developed an improved sampling algorithm and an energy model for protein loop prediction that yields a smaller root mean square deviation from the native structure. They discussed their results in the context of the accuracy of continuum solvation models. Xiang (9) discussed the advances in protein homology modeling and the contribution of loop structure predictions for the overall prediction of protein structures.

According to Radivojac *et al.* (10), some of the intrinsically disordered regions of protein structures consist of long loop regions with functional roles. Since conformation and dynamics are intrinsically related, any improvements in understanding one will likely improve the understanding of the other. Thus, the methods proposed in this paper may help us to better understand the relationship between functional motions and loop conformations.

Protein loops play an important role in protein function since they are often exposed to the solvent environment and hence may readily interact with other molecules. It is widely understood that their structures are not random coils (even for longer loops), and thus have some defined characteristics (11). Conformations of loops play a significant role in protein docking (12) and in stabilizing active sites through loop-scaffold interactions (13). Smith *et al.* (12,13,14) investigated the idea of guiding protein-protein interactions through contacts between surface loops in proteins. Hence flexibility of protein loops and their dynamics are important factors for understanding protein functions, as demonstrated further by Yao *et al.* (15) who used sampling algorithms to explore conformations of flexible loops. Krieger *et al.* (16) have shown that folding mechanisms in proteins vary widely depending on native-state topology and details such as the relative contact order (RCO). This indicates that protein loops and their topologies might also play an important part in protein folding. Conformational evaluation of loops and their structural variability was studied by Li *et al.* (17) who indicated the importance of loop structures for protein design.

Hu *et al.* (18), demonstrated through high-resolution design of protein loops that small changes in protein energetics can perturb the structure of proteins. They studied longer loops adopting specific conformations with the Rosetta molecular modeling program to find low-

energy sequence-structure pairs. Their results suggest that the high-resolution design of protein loops may become feasible.

We previously investigated (19) the fluctuation dynamics of the tubulin dimer to elucidate the functional motions that might relate to activities such as binding, polymerization and assembly and discovered that a loop that covers the GTP binding site moves in coordination with a large-scale rotation between the α and β subunits. This illustrated how loop motions can be controlled by the large domain motions and can be slow. Also we investigated (30) the enzyme triose phosphate isomerase and observed that its binding site loop opened and closed only in the intact dimer, and not in the monomer, in a slow motion coordinated with a large-scale domain-domain motion.

Espadaler *et al.* (20) developed ArchDB, an automated classification tool for the structures of protein loops that connect different supersecondary structures and play an important role in initiating and maintaining the overall functions of a protein. Oliva *et al.* (20,21), computationally derived an extensive characterization of loop conformations that could enhance model building by comparison studies. Groban *et al.* (22) illustrated phosphorylation driven changes in loop conformation using the activation loop in CDK2. Kolodny *et al.* (23) approached the 'loop closure problem' using inverse kinematics. They proposed an algorithm for generating conformations of candidate loops within gaps in protein structures to complete protein structures so that their biological functions can be determined. Gerstein and Chothia (24) demonstrated the significant mobility of surface loops that can move over a distance of 10 Å to cover the active site, and showed that this motion is propagated outwards towards other regions of protein structure that have no contact with the ligand. They suggested that the whole protein consists of several different shells of increasing mobility. Andrec *et al.* (25) have developed a novel

approach for detecting statistically significant differences in the structures of loops between crystal and NMR-determined structures. Their approach is based on structural superposition and the analysis of the distributions of atomic positions relative to a mean structure. Their studies indicate that physical factors and the environment play a role in determining protein conformations. Sudarsanan *et al.* (26) used information from the backbone conformation of dimers to develop an automated method for modeling the backbones of protein loops that obtains near-native loop conformations from an ensemble of sterically allowed conformations. Street *et al*. (27) investigated the physical-chemical determinants of the turn conformations in globular proteins, concluding, as have many others, that turns can be classified into a small number of discrete conformations. Kempf *et al*. (28) examined how the loops in triosephosphate isomerase facilitate substrate access and catalysis. They investigated the dynamic requirements for functional hinges and elucidated the important principle of motional freedom and flexibility requirements for active-site loops, which control the open and closed states of active sites. Their results demonstrate the importance of catalytic hinge design in proteins.

In the present study we will investigate loop motions with elastic network models. We are interested in analyzing loop motions to see if they move independently or in coordination with large domain motions. We thus are able to identify the local motions of loops that make the largest contribution to the overall domain motions. The focus is on the dynamics of all the surface loops present in five diverse proteins: reverse transcriptase, triosephosphate isomerase, tubulin, protease, and myoglobin. Each of these proteins is distinct from the others in its topology and function, thus providing a small but diverse test set. The loops present in these proteins are known to have diverse functional behaviors. The choice of reverse transcriptase was based upon the importance of the loop motions that provide access to the polymerase site,

specifically related to how the fingers and the thumbs move to open and close this site, as shown by Bahar *et al.* (29). Including triosephosphate isomerase was motivated by our previous observation of the importance of the loop moving over the active binding site (30). The loop covering the GTP binding site in tubulin was also shown previously to be coordinated with a slow motion of the protein (19). We have previously shown that this motion occurs together with the dominant motion, which is a rotation between the two subunits. The flaps of the protease are well known loops that regulate access to its binding site. The behaviors of the loops in these five proteins are examined in detail below and our findings suggest that functional loops behave in coordinated ways with the rest of the structure, rather than as random motions.

C.2 Materials and methods

Normal Mode Analysis

To study the kinematics of residues constituting loops we use **NMA** (normal mode analysis) on the coarse-grained elastic network models of structures. The structures are represented by C^{α} atom coordinates only. Harmonic springs are used to connect the C^{α} atoms in order to represent the protein structure as an *elastic network*. The Gaussian Network Model (**GNM**) is one of the simplest of elastic network models, originally applied to protein dynamics by Bahar *et al.* (31) and Haliloglu *et al.* (31,32) who applied the approach of Tirion *et al.* (33) in a coarse-grained way to both bonded and non-bonded contacts in proteins and represent their interactions with a single universal spring parameter. This model has its deep origins in the rubberlike elasticity theory of Flory, James and Guth, James and Guth, Kloczkowski *et al.*, Skliros *et al.* (34,35,36,37,38). Each normal mode corresponds to a different frequency of oscillation. Extensive applications of NMA to biological and chemical systems have been discussed in Cui *et*

al., Jernigan and Kloczkowski, Sen *et al.* (39,40,41). The Anisotropic Network Model (ANM) developed by Atilgan *et al.* (42), can be used to compute the *directions of motions* of all points in the structure with the coarse-grained elastic network model, whereas the original GNM provided only the amplitudes of motion. We employ the ANM model throughout our following analyses.

Kinematics of Proteins

Our method of solving the kinematics of proteins in the coarse-grained representation is based on Lagrange's equation for the potential and kinetic energy of the system, as described by Kim *et al.*, Kim *et al.* a, Kim *et al.* b, Schuyler and Chirikjian, Schuyler and Chirikjian (43,44,45,46,47). As a first step, a rigid body translation and rotation of the structure is performed, so that the center of mass lies at the origin of the coordinate system and the moment of inertia tensor is diagonal. This procedure is described in detail in Supporting Information Section A.

To solve for the kinematics of the protein we define the coordinates as

$$\begin{aligned} &\vec{R}_i(t) = \vec{R}_i(0) + \overrightarrow{\Delta R}_i(t) \Rightarrow \overrightarrow{\Delta R}_i(t) = \vec{R}_i(t) - \vec{R}_i(0) \\ &\overrightarrow{\Delta R}_i(t) = \left[\Delta x_i(t), \quad \Delta y_i(t), \quad \Delta z_i(t)\right] \end{aligned} \tag{1}$$

where $\vec{R}_i(t)$ and $\vec{R}_i(0)$ are the instantaneous and the starting position vectors for the i^{th} point and $\overrightarrow{\Delta R}_i(t)$ is the displacement vector. The potential energy of the system can be written as (details shown in Supporting Information Section B)

$$V = \frac{1}{2}\Delta\vec{R}^{T}(t) K \Delta\vec{R}(t)$$

where K is the matrix of the order $3N \times 3N$ which depends on spring constants and initial position vectors of all points in a structure.

If we take $\Delta\vec{R}(t) = \left[\Delta R_1(t) \quad ... \quad \Delta R_N(t)\right]$ for each time *t* then we find the solution for the elastic model for all values of *i* is given by

$$\Delta\vec{R}(t) = \sum_{i=1}^{3N} \left[\frac{1}{\sqrt{\lambda_i}} \sin\left(t\sqrt{\lambda_i}\right) e_i e_i^T \Delta\vec{R}(0) + \cos\left(t\sqrt{\lambda_i}\right) e_i e_i^T \Delta\vec{R}(0)\right] \tag{2}$$

where λ_i and $\sqrt{e_i}$ $(i = 1,...,3N)$ are the $\sqrt{}$ eigenvalues (square of angular frequencies) and eigenvectors (normal modes) of the system. For more details see Supporting Information Section C. For evaluation of the motions of loops we select the components of the 3*N*-dimensional vector $\Delta\vec{R}(t)$ that correspond to the coordinates of the residues in the loop. We then study their time evolution by solving Eq. (2).

Identifying the dominant normal modes by Fourier Analysis

The essence of equation (2) is that it calculates the displacement of each coordinate from the equilibrium position at any given time *t*. We set the initial conditions of the fluctuations in such a way that the initial moment of inertia of the system is zero. We want information from time-dependent displacements to reconstruct the signal. The solution comes from the Nyquist-Shannon theorem, Shannon (48), which states that if the signal *x(t)* has no angular frequencies higher than Ω_0, it is completely determined by giving the ordinates as a series of points separated by time intervals $\frac{\pi}{\Omega_0}$. For the current case, we know that the maximum angular frequency of the system is the square root of the maximum eigenvalue, $\Omega_0 = \sqrt{\lambda_{\max}}$. This corresponds to selecting a sampling period of $T_s \le \frac{2\pi}{2\Omega_0}$ or $\Omega_s \ge 2\Omega_0$. Furthermore we see from Eq. (2) that the

motions of residues can be expressed as a combination of sinusoidal functions, making them periodic.

The maximum period of the system that defines the final time in our calculations is $T_{\max} = \frac{2\pi}{\sqrt{\lambda_{\min}}}$, $\lambda_{\min} \neq 0$. For each element of $\Delta\vec{R}(t)$ we calculate its time evolution, following from Eq. (2), at time intervals $t = 0, T_s, 2T_s, ..., sT_s$, with $s = \left\lceil \frac{T_{\max}}{T_s} \right\rceil$ up to $T_{\max}$. To each of the $3N$ coordinates we can assign an s-length discrete time signal, called $H(n) = \Delta R(nT_s)$ that is periodic with the period $T_{\max}$.

The Discrete Fourier Transform DFT of this signal is given by:

$$F_H(k) = \sum_{n=0}^{s-1} H(n) e^{-2\pi jkn/s}, \quad k = 0, ..., s-1 \tag{3}$$

To calculate all the s-entries of that signal we require s^2 multiplications and $s(s-1)$ additions. The Fast Fourier Transform (FFT), as proposed by Cooley and Tukey and Singleton (49,50), significantly reduces the computational cost.

In order to recover $H(n)$ from $F_H(k)$, we apply the inverse Discrete Fourier Transform defined as:

$$H(n) = \sum_{k=0}^{s-1} F_H(k) e^{2\pi jkn/s}, n = 0, ..., s-1. \tag{4}$$

FFT also applies to the inverse Discrete Fourier Transform, and the interested reader might refer to the Digital Signal Processing literature such as given in Antoniou, ElAli, Hayes (51,52,53). Eqs (9-10 in (52)), imply that the magnitude of $F_H(k)$ is symmetrical about the

point $k=\frac{s}{2}$, thus $\left|F_H(k)\right|=\left|F_H(s-k)\right|$, and $\angle F_H(k)=-\angle F_H(s-k)$ for the phase of the signal. The lowest frequencies of the signal are located at the ends of $F_H(k)$ whereas the highest frequencies are located in the middle. It is a symmetric signal with $F_H(p-K/2)=F_H(p+K/2)$, where $k,p\in[1,2,...,K]$. It was also noted in reference (52) that the distances between the successive values of k in $F_H(k)$ are given by the angular frequency resolution $\frac{\Omega_s}{s}$. The correspondence between indices k of $F_H(k)$ and the eigenvalues of the system can be specified as given in Supporting Information Section D.

To evaluate the impact of the lowest frequency motions on the system, we first identify the proper k indices in $F_H(k)$. Then we set the value of $F_H(k)$ for the k's that do not belong in that range to zero which leads to the new FFT $F'_H(k)$. Then we take the inverse DFT (Discrete Fourier Transform) of $F'_H(k)$, thus obtaining a new discrete time signal $H'(n)$ that depends only on the lowest normal modes of the system. Finally we compute the Pearson correlation between $H'(n)$ and $H(n)$ (46). The higher the value of the correlation, the greater the impact of the lowest normal modes is on the motions of the system.

Computing Changes in Internal Distances

We also consider the changes in the internal locations of the structure points with ANM. This is the change in internal distance, computed as

$$<(\Delta R_i-\Delta R_j)^2>=<\Delta R_i^2>+<\Delta R_j^2>-2<\Delta R_i\cdot\Delta R_j> \quad (5)$$

These values are obtained directly from the inverse of the Hessian matrix from which the normal modes are derived.

$$< (\Delta R_i - \Delta R_j)^2 > = \Gamma^{-1}{}_{ii} + \Gamma^{-1}{}_{jj} - 2\Gamma^{-1}{}_{ij} \quad (6)$$

where Γ is the matrix of second derivatives of the potential energy (Hessian) for the structure for which the normal modes (e_i) are computed. Since there are six zero eigenvalues in ANM corresponding to the rigid body motions, Γ is not invertible. Thus, instead we compute its pseudo-inverse: $\Gamma^{-1} = \sum_{i=7}^{3N} \frac{1}{\lambda_i} e_i e_i^T$.

Loop Identification

In our study we first identify the surface loops on the proteins. We identify these loops in proteins by excluding all residues identified by DSSP (54) as H, G, I, or E, corresponding to standard α, 3_{10}, and π-helices, and β-strands, respectively. We retain isolated beta-bridges and hydrogen bonded turns to prevent short loops interrupted by these elements from being discarded. We focus on surface loops by also rejecting any residue having surface exposure less than 5% in an extended A-X-A chain using NACCESS program (55) to compute relative solvent accessibility. We also set the requirement that the length of a loop must be four or more residues. Visual inspection of the 5 protein structures studied in this work show that this selection appears reasonable. The identity of all loops studied here is given in the Supporting Information Section E. The functional loops are defined as those loops which contain one or more functional sites. Information about the functional parts of the proteins and the functional loops is provided in the Supporting Information Section F. Functional information is derived from the NCBI Protein database and manually related to the corresponding protein structures.

C.3 Results and Discussion

The main purpose of this study is to answer three major questions. (i) Do protein residues move overall independently, or do they move in coordination with the entire structure? (ii) Do

loops in proteins move along with the whole structure or do they exhibit a different behavior? (iii) Do the functional loops move independently or in coordination with the slow motions, and do they move like non-functional loops? We attempt to answer these questions by investigating five different proteins in terms of function and topology, which are given in Table 1. To address the first question, that is whether proteins residues move individually or collectively we employ the Anisotropic Network Model (ANM). This model depends on the whole structure of the protein through a connectivity matrix, dependent on topology. In this study, we find that there is a close correspondence between the behavior of loop motions and the entire structure as observed in the protein reverse transcriptase (19). In Figure 1, we show the correlation between the motions obtained for the first 6 normal modes (the slowest motions) for all residues in comparison with the loop residues for four proteins studied in this work. (For myoglobin see Supporting Information G.) Residue indices are sorted according to the increasing values of correlations. The first six modes are the slow, collective, low frequency motions of the structure. We see that in this case, the overall motions of the loops do not differ much from the motions of the whole structure.

Table 1. The proteins used in this study

Name	PDB ID	State	Residues	# Loops
Tubulin	1TUB	Heterodimer	867	36
HIV-1 reverse transcriptase	1DLO	Heterodimer	971	47
Triosephosphate isomerase	1WYI	Homodimer	496	20
Protease	1J71	Monomer	338	18
Myoglobin	2V1K	Monomer	153	5

For each coordinate of each residue, we calculate the displacement from the initial position at several time instances. We thus construct a discrete time signal *H*. The details of how we obtain this signal are explained in the Methods section. In the computations given below, *H* is the kinematic response of each coordinate of each residue based on all normal modes whereas *H'* is the similar kinematic response based on a subset of the lowest normal modes. Thus *H* is the full FFT signal while *H'* is the FFT signal corresponding to the low frequency normal modes. High correlations between these two will indicate a dominance of the low frequency motions. In Figure 1 these correlations are shown, and we can see that when all protein residues are considered that they move with the global (collective) domain motions since the percentage of motion represented by the six lowest normal modes is always above 50%.

We have also computed the mean correlations $< \rho_{H,H'} >$ between *H* and *H'* averaged over the residues within the loops, for all loops in a given protein. Similarly, we compare correlations computed by using all normal modes in Eq. (2) with those obtained by using only the slowest modes (details are given in the Methods section). Results are shown in Figure 2 for loops belonging to chains A and B of reverse transcriptase and of tubulin (for triosephosphate isomerase, protease and myoglobin see Supporting Information G). Similarly as in Figure 1, the loop indices are sorted according to ascending values of the correlations. Circles identify functionally important loops.

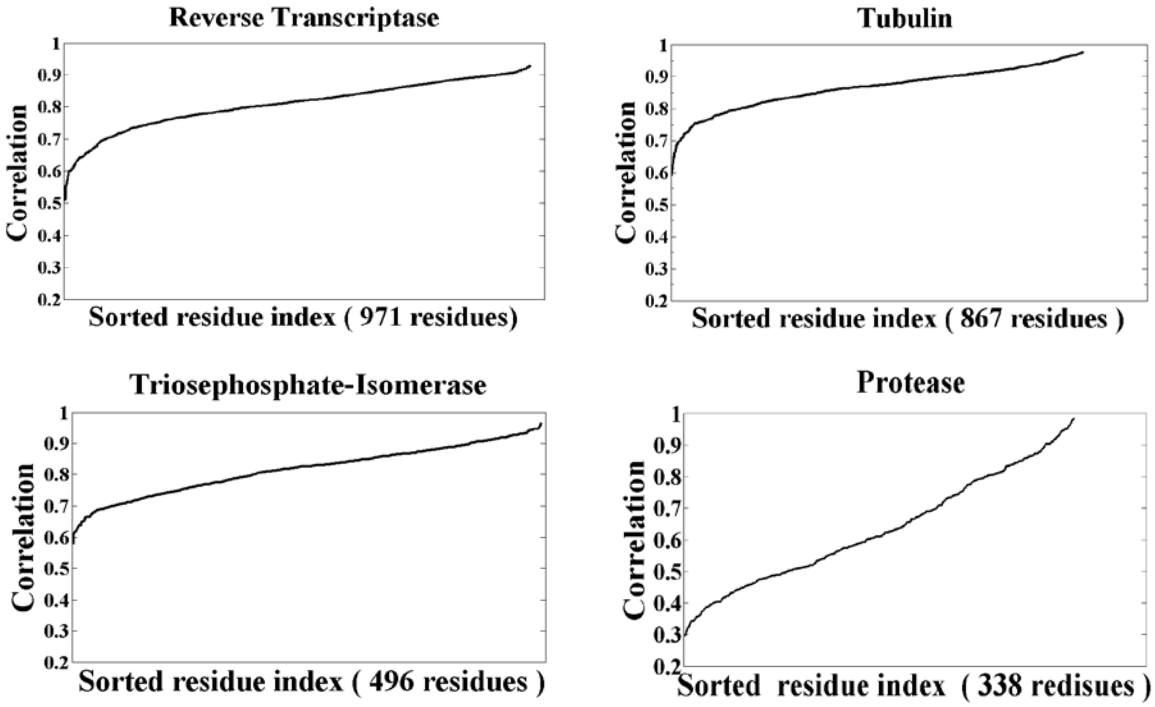


Figure 1: Impact of the motions of the first 6 normal modes on the overall motion, for all residues of reverse transcriptase, tubulin, triosephosphate-isomerase and protease (myoglobin in Supplemental Information G).

From Figure 2 we see that the mean impact of the first 6 normal modes on the loops ranges between 65-99%, a slightly larger range of correlations than the impact of the first 6 normal modes on all residues of the proteins. Thus from Figure 2 we conclude that protein loops move as a part of a domain to a somewhat greater extent than all other parts of the protein structures. The slightly larger correlations may be attributed simply to the loops being investigated residing on the outside of the structures

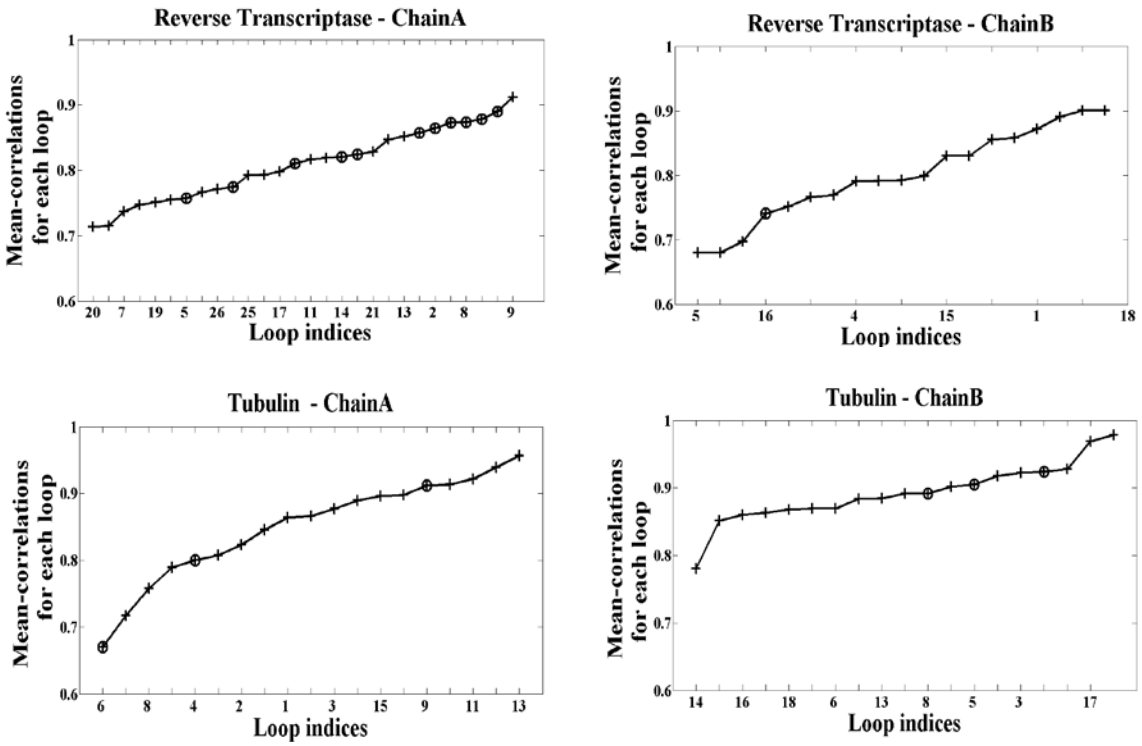


Figure 2: Mean correlations of the motions derived from the first six normal modes with the total motions for each of the loops of reverse transcriptase and tubulin. The functional loops are denoted by circles.

For the loops of reverse transcriptase we know that loops with indices 2,4,5,6,8,14,18,22,24,27,28 on chain A and 8 on chain B control access to the catalytic residues or contain binding residues. We see from Figure 2 that motions of these functional loops do not differ significantly from motions of other, non-functional loops. Likewise for tubulin (Figure 2), the loops with index numbers: 4, 6, 9 from chain A and 5, 8, 9 from chain B responsible for regulation of the interactions with other tubulin dimers do not differ in behavior much from the average behavior of the loops of tubulin. We also randomly generated 100,000 partitions of the loops into two groups (data not shown) where the smaller group was the minimum of 15 or half of the loops. The most significant partitions (determined by the amount of difference in average mean squared fluctuation) were either trivially different from one another or corresponded to groups of loops that were farthest from the protein's center of mass.

In Supporting Information F we show the location of functional loops on each structure. Hence, the answer to the third question could be that functional loops do not move in a more coordinated way with the rest of the structure than regular loops, although some individual loops may do so (see Fig. 2).

Normal mode calculations are often performed to elucidate which residues or atoms in a molecular structure are the most mobile. Active site residues are often held relatively rigid. Two supporting cases here are reverse transcriptase and protease where the catalytic residues are within a cleft where they are held relatively rigid. It is the movement of the surrounding structural elements that regulate access to these catalytic residues that facilitate access to the protein active site. However, another quantity which may be informative is the internal mean square distance changes described by Eq. 5. Internal mean square distance changes can be calculated directly from the Hessian matrix used to generate the normal modes in ANM using Eq. 6. We have employed (57) ANM models built with uniform springs with cutoffs ranging from 10-15 Å, as well as with springs having inverse square dependences on distance and obtained similar results. The mean square internal distance fluctuations, $< (\Delta R_i - \Delta R_j)^2 >$, describe the change within a structure; how the normal modes stretch, compress, or otherwise rearrange the internal structure locally. If this change in internal distance is zero for a given (i, j) pair, it means that the two points move together fully rigidly (the distance between them does not change). We have analyzed the present five protein structures (data not shown) and concluded that the areas of a protein with the smallest internal mean square distance changes are the cores, and as one moves further away from the stable cores the internal distance fluctuations increase. Figure 3 shows the mean internal RMSD for each loop of reverse transcriptase and tubulin. (For triosephosphate isomerase, protease and myoglobin see Supporting Information G). We see that

those loops that are functional do not have lower or higher RMSDs than nonfunctional loops. Hence the nonfunctional loops do not differ in the internal conformational behavior from the nonfunctional ones.

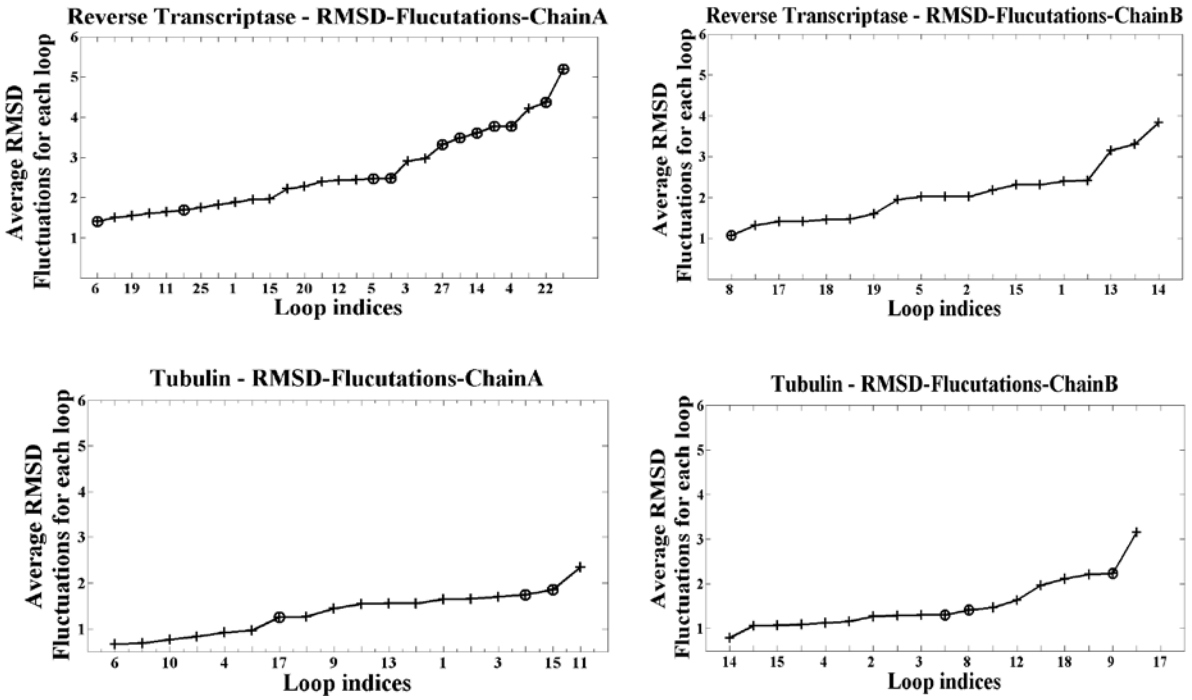


Figure 3: Mean RMSD for the loops of chain A and chain B of the reverse transcriptase and chain A and chain B of tubulin. Functional loops are indicated by circles.

It is also of interest to locate the loops on protein structures that have the highest correlations according to Figure 2. These loops are highlighted in Figure 3 for the HIV-1 reverse transcriptase structure and are mostly associated with the areas surrounding catalytic residues (see Supplemental Section H for the other four proteins). In Figure 4 we show a zoomed in view of the polymerase active site where a large loop hangs over the opening between the thumb and fingers. This loop may act to regulate substrate access to the catalytic residues and influence binding on the interior of the thumb and fingers (white arrow in the lower part of Figure 4) domains. Yellow surfaces correspond to experimentally verified nucleotide binding residues. Another loop with a very high correlation coefficient is marked with a solid black arrow in part C that may also interact with bound substrate. It is likely that many of these loops owe their high

correlation to the large hinge motion through the middle of the structure that is seen in the dominant mode of motion.

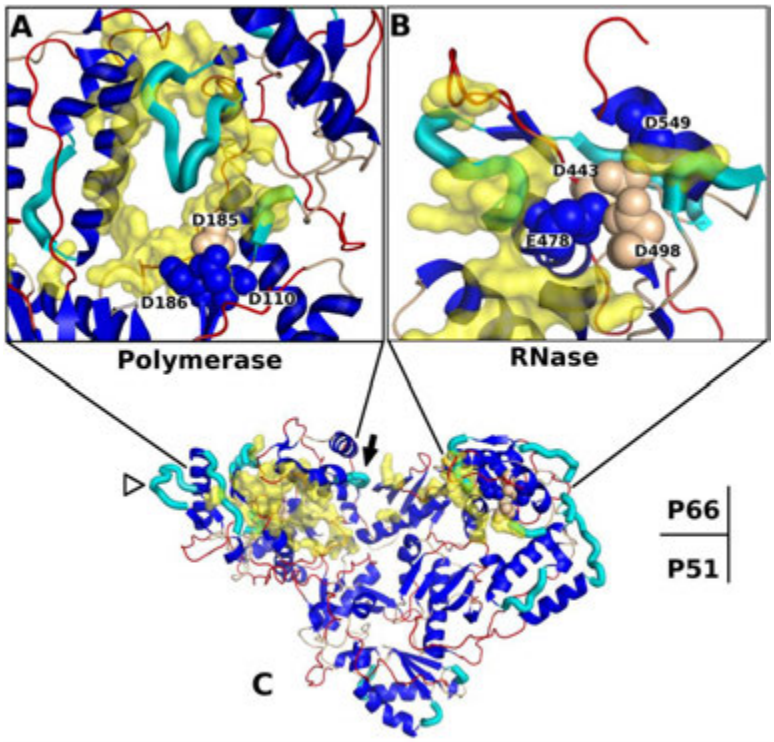
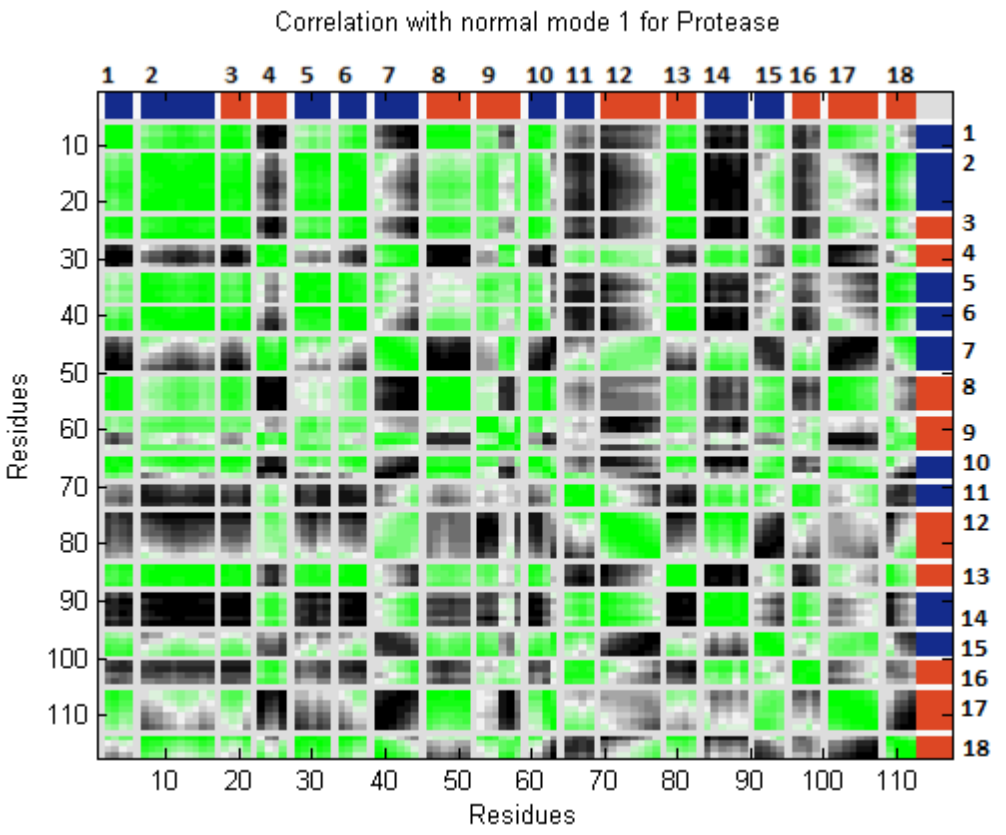


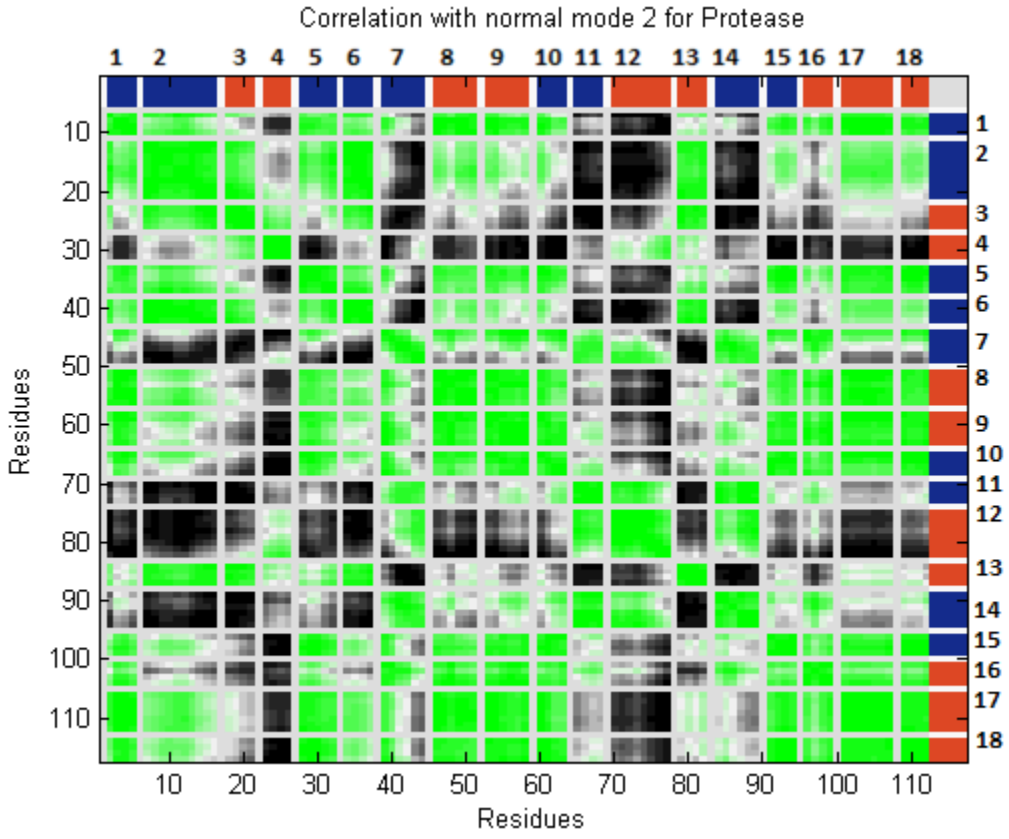
Figure 4: We highlight with thick cyan tubes the surface loops of HIV-1 reverse transcriptase that have the highest average correlation (coefficient > 0.87) of motion between the first six modes and all modes. Catalytic residues of the polymerase and RNase domains are labeled and shown as spheres. Yellow molecular surfaces are shown for residues experimentally determined to contact the nucleotide template. The other surface loops (see Methods) are colored red with other loops in tan. (A) Zoomed and rotated to show the polymerase catalytic domain with the fingers on the right and thumb on the left. (B) RNase catalytic domain. (C) The P66 and P51 dimer is shown. The white arrow head points to the polymerase finger domain which contains three cyan loops. The filled black arrow points to a loop which is likely to interact with the nucleotide chain in the dominant modes of motion. See Supplemental Information H for similar figures for the other four proteins.

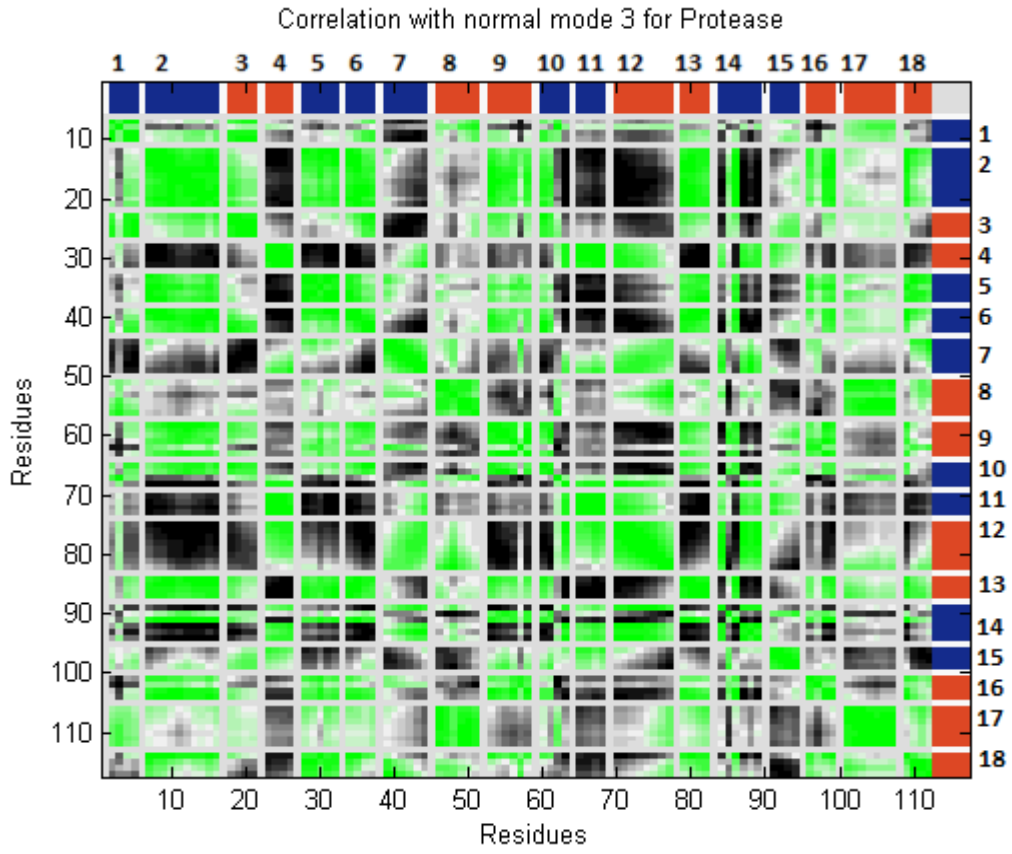
For the sake of finding the amount of correlation or anti-correlation among the residues which correspond to only functional and non functional loops, we have reduced the correlation map for ANM from all residue set up to loop residues only. We have analyzed these correlation maps for all these functional and non-functional loops for all five proteins for only first six normal modes which correspond to the global motions of these proteins and as shown in figure 5

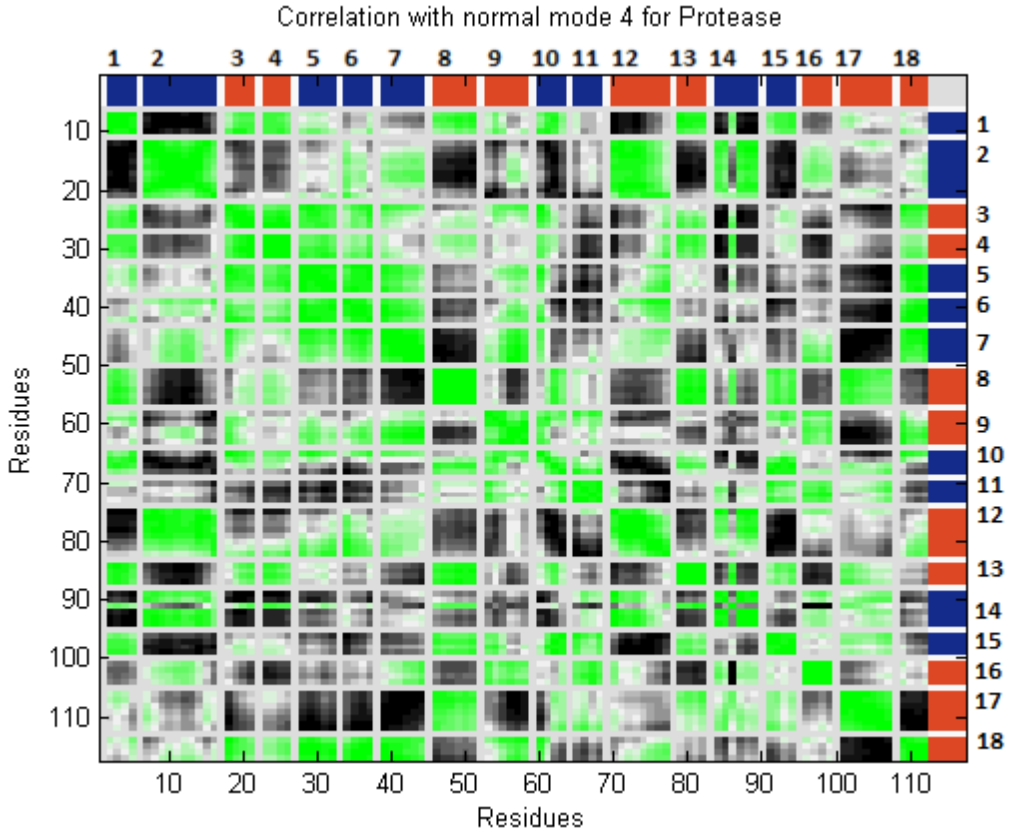
(for protease) and also in Supporting Information I (for myoglobin , triosephosphate isomerase , tubulin and reverse transcriptase) , these correlation plots for only loop residues exhibit a significant amount of correlation among mainly the functional loops . Also in some cases, there is an extent of anti-correlation among certain functional loops in particular modes which again explains a particular functionality for that protein .

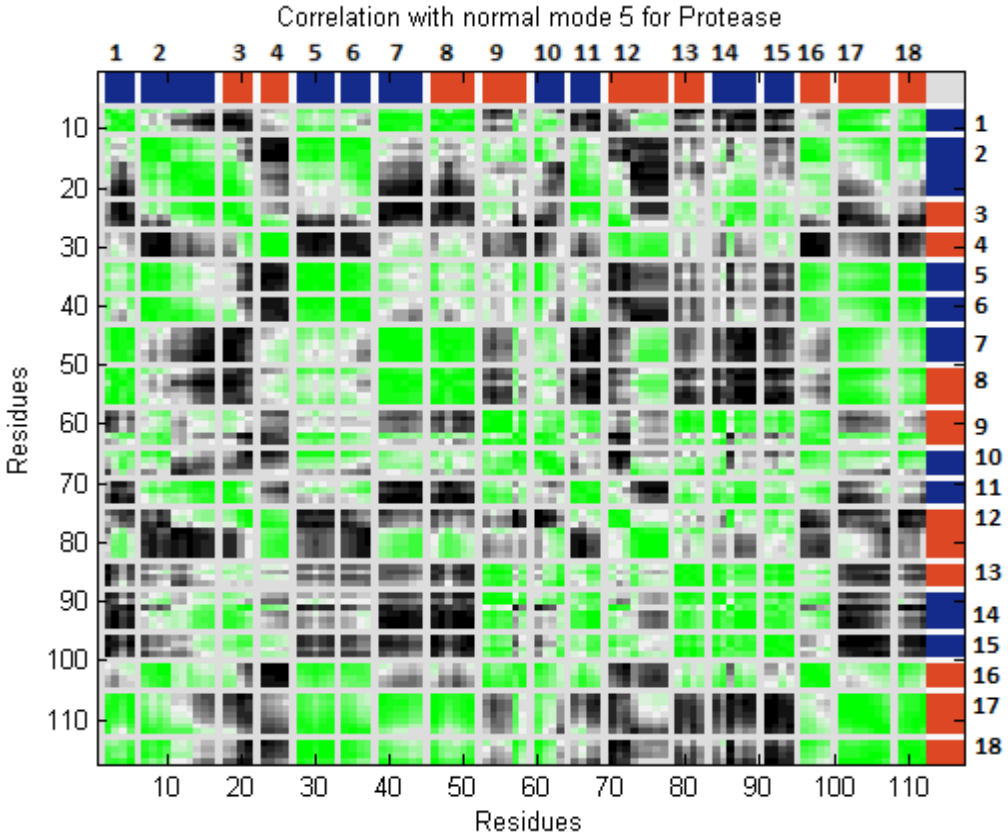
In figure (5), the functional and non-functional loops are shown in blue and red respectively. The total number of loops of protease is 18 of which only 9 (loop indices 1,2,5,6,7,10,11,14 & 15) are functional loops . For protease, total number of residues is 338 , but as we have excluded the residues which belong to other secondary structures like alpha helix or beta sheet and only included the residues belonging to the loops, the number of residues is much lower than 338 . We have used a white demarcation line between the two adjacent loops to clearly distinguish the correlations and anti-correlations in these loop residues. The correlation ranges from +1 (shown in green) to -1 (black) . Also there has been cases where we have found no correlation (or correlation = 0) which has been denoted by white color.

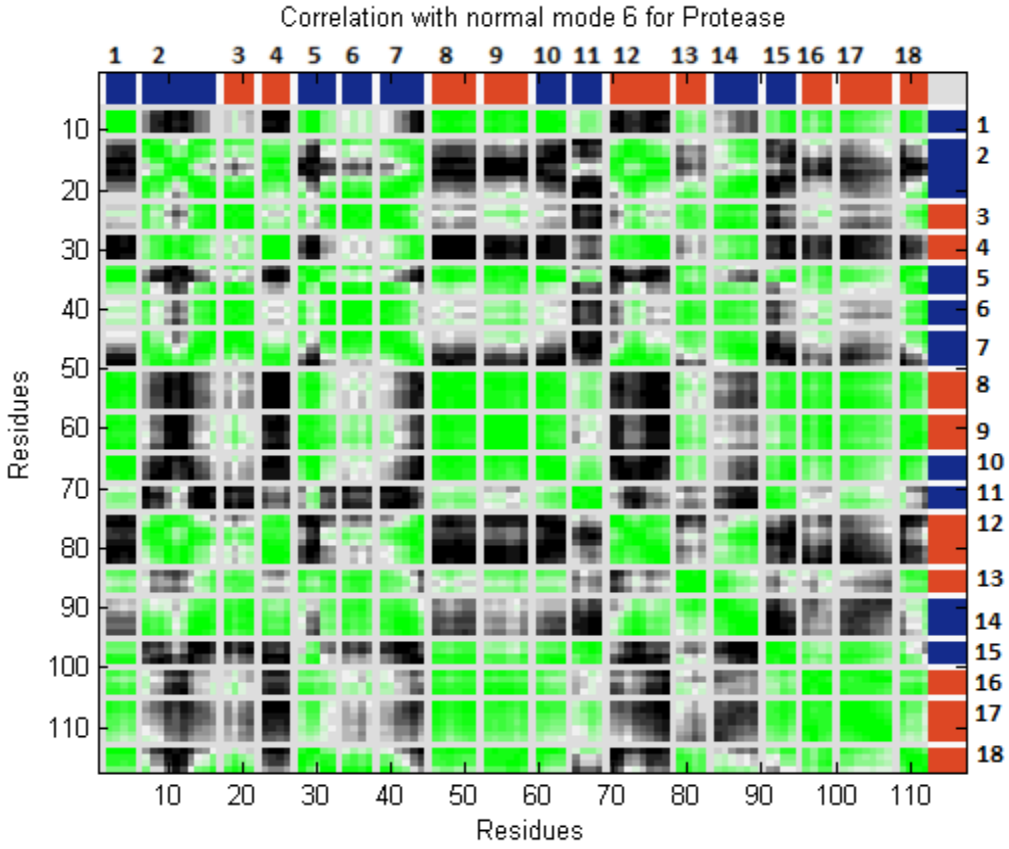

Correlation with normal mode 1 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
10 20 30 40 50 60 70 80 90 100 110
Residues
Residues


Correlation with normal mode 2 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
10 20 30 40 50 60 70 80 90 100 110
Residues
Residues


Correlation with normal mode 3 for Protease
Residues
Residues
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
10 20 30 40 50 60 70 80 90 100 110


Correlation with normal mode 4 for Protease
Residues
Residues
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
10 20 30 40 50 60 70 80 90 100 110


Correlation with normal mode 5 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
10 20 30 40 50 60 70 80 90 100 110
Residues
Residues
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18


Correlation with normal mode 6 for Protease
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
10 20 30 40 50 60 70 80 90 100 110
Residues
Residues
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

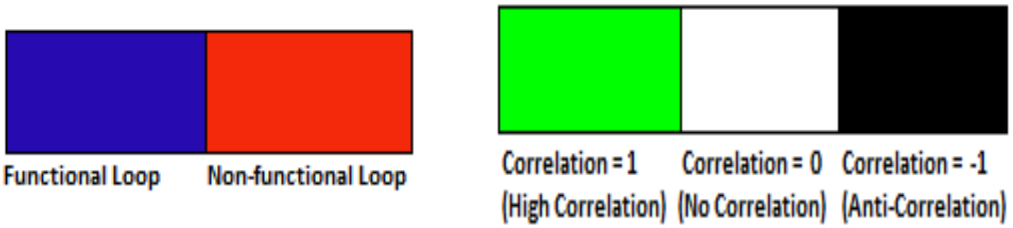
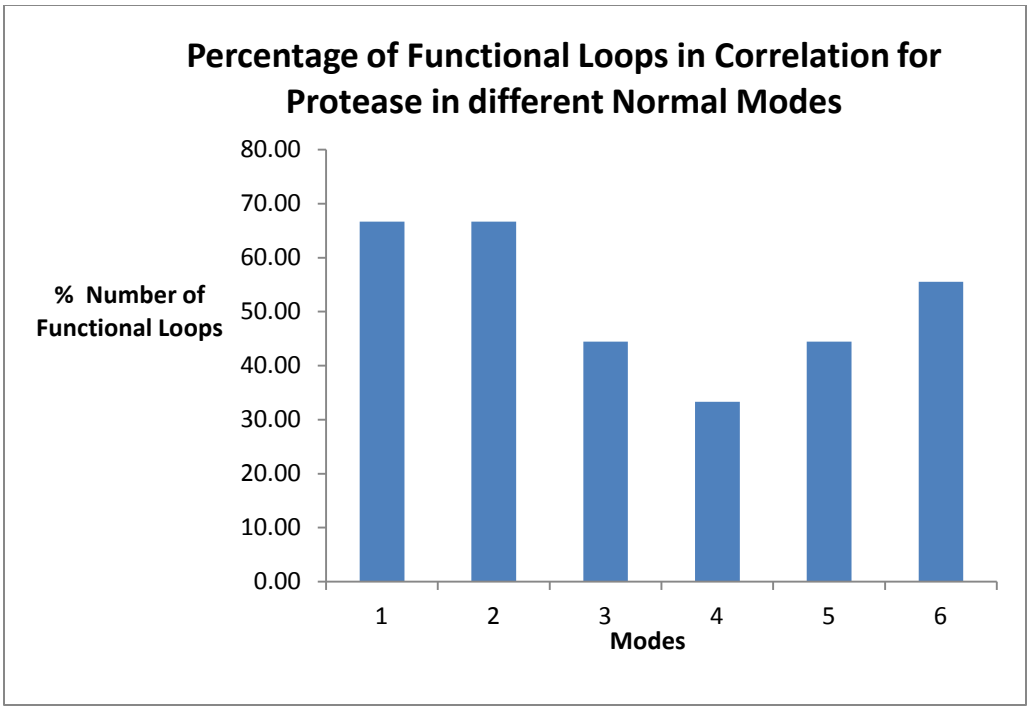
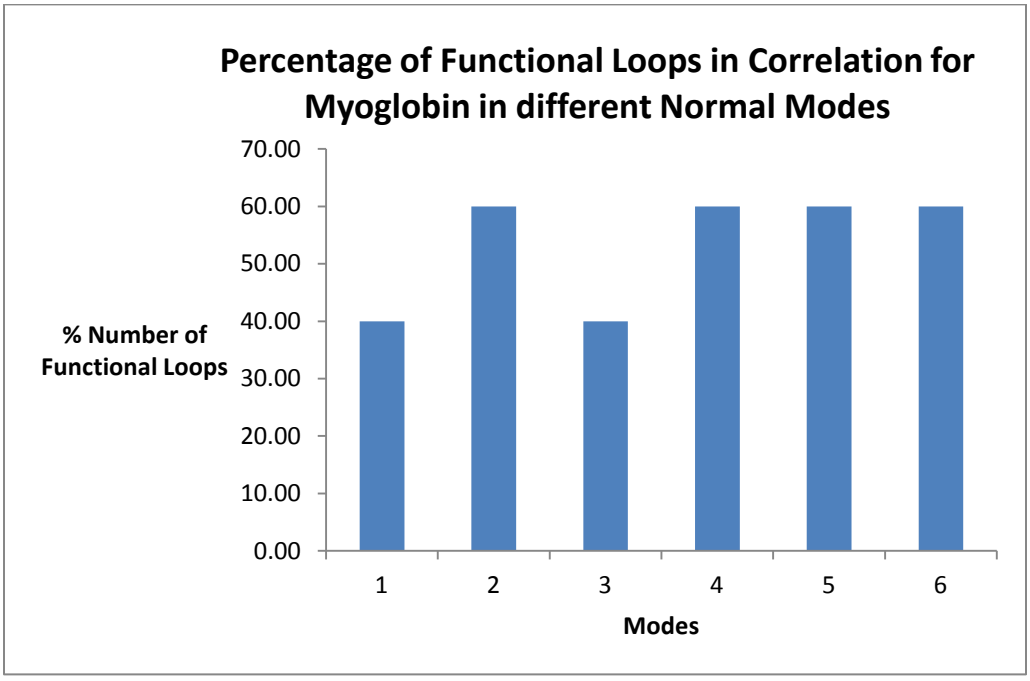
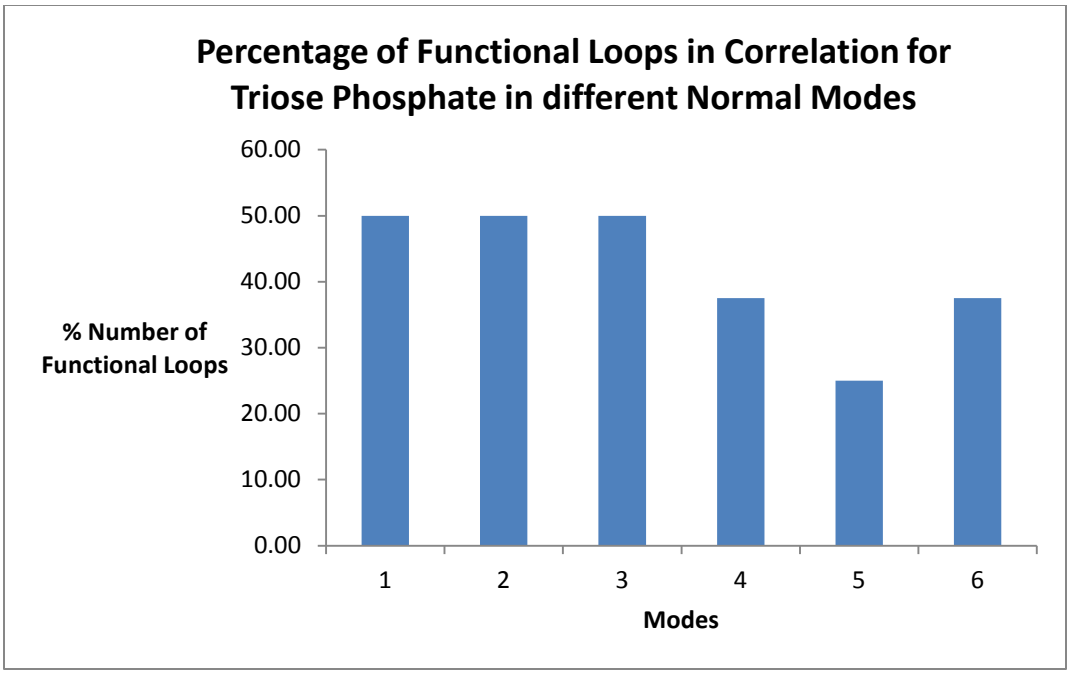


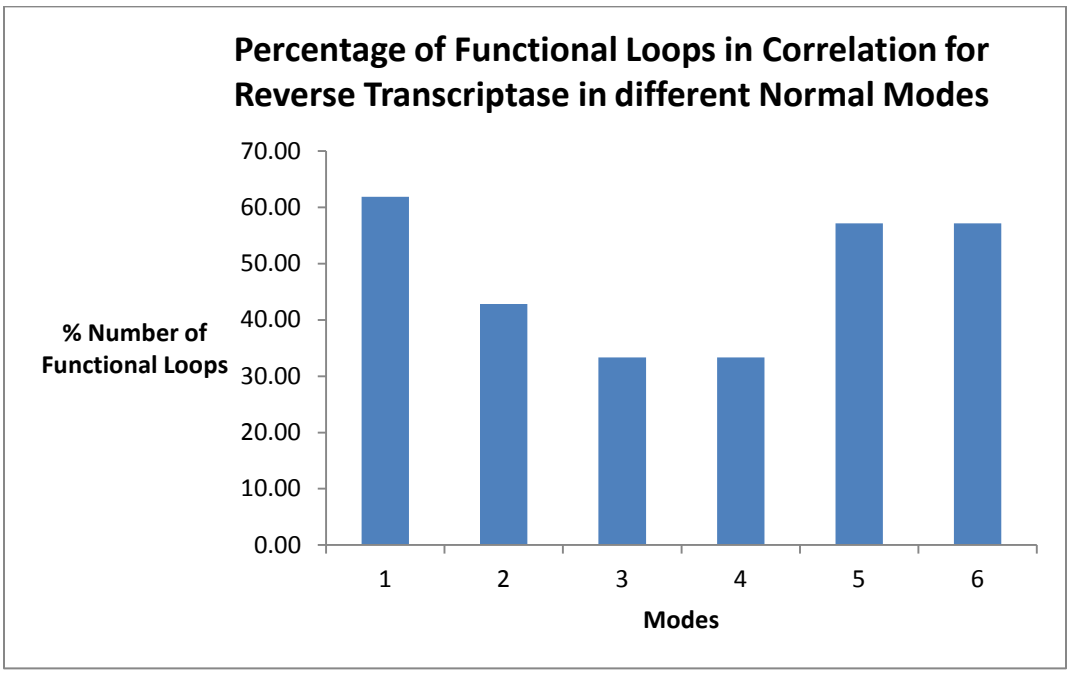
Figure 5: Correlation Plot for first six normal modes for the functional and non-functional loops of protease. The functional loops are marked in blue, while non-functional ones are shown in red color. We use green, white and black colors to show the transition from the highest correlation (green) to the lowest (black) . As shown in the diagrams, numbers (1 to 18) indicate loop indices for protease.

We have also calculated the percentage of total number of functional loops which move in correlation for all these five proteins under the first six normal modes. This is to address in a more informative way to the third question, i.e. whether the functional loops behave in a more coordinated way with the slow motions or if their behavior is independent of the global motion. Figure 6 shows the results for all the five proteins. We have found approximately 40% to 70% of the total number of functional loops move in coordination for majority of these proteins in most of the different lowest six normal modes which is again significant considering that these functional loops are not adjacent and some of them are really far apart from each other.


Percentage of Functional Loops in Correlation for Protease in different Normal Modes
% Number of Functional Loops
80.00
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes


Percentage of Functional Loops in Correlation for Myoglobin in different Normal Modes
% Number of Functional Loops
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes


Percentage of Functional Loops in Correlation for Triose Phosphate in different Normal Modes
% Number of Functional Loops
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes


Percentage of Functional Loops in Correlation for Reverse Transcriptase in different Normal Modes
% Number of Functional Loops
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
1
2
3
4
5
6
Modes

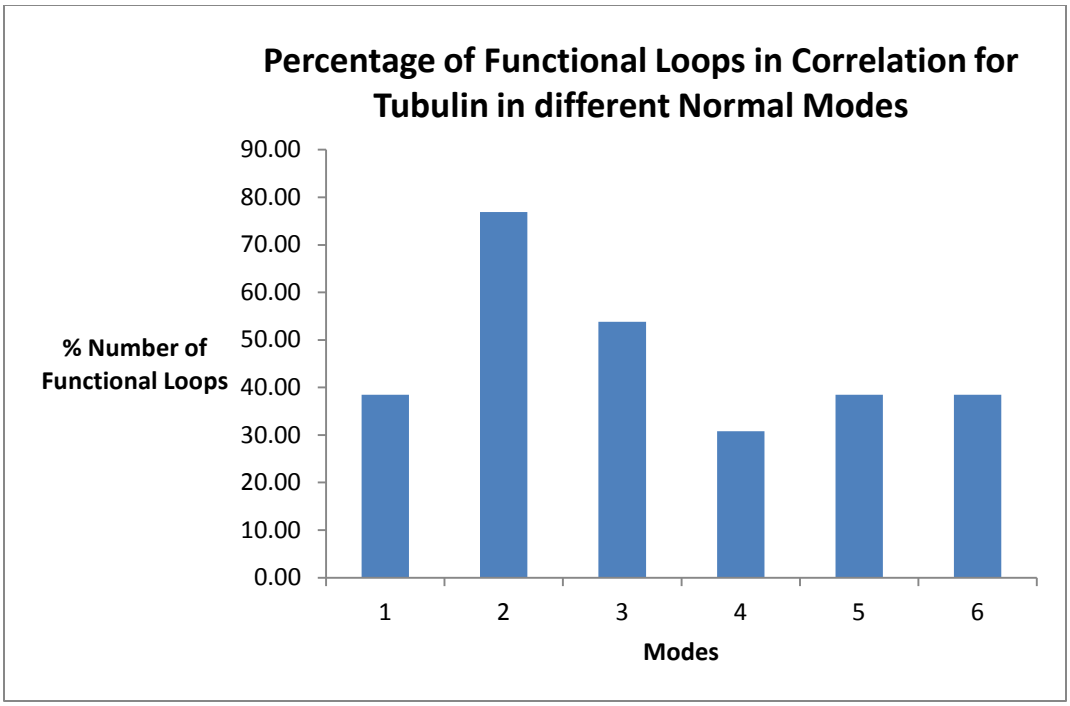


Figure 6: Percentage number of Functional Loops moving in correlation for five proteins under first six normal modes.

C.4 Conclusion and Outlook

In the present paper, we have considered the motions of the surface loops in five proteins. By applying a novel method that combines ANM and FFT we are able to identify which normal modes have the largest impact on the motions of individual loops. We observe a broad range of behaviors, with some loops moving in the slowest modes, which implies that their motion is strongly linked with the global, collective motions of the structure and others moving with the fast modes.

We know that loops are parts of protein structure that are likely to be more susceptible to the influence of the external environment. Environmentally influenced changes in loop structures or dynamics may lead to radical structural changes of the whole protein. The reverse may hold also because of the protein's cohesiveness, so that external influences changing loop

conformations could also push the large domains into different positions, leading to allosteric transmission.

Prediction of loop motions with and without external environmental influences can lead to a better understanding of the functions of loops and their mechanisms. More specifically, we can potentially identify the mechanics of the hyper-variable loops of antibodies and how they may move in response to the presence of a specific antigen. We can also try to understand the mechanism of motions at the polymerase sites, the mechanism of GTP binding sites in tubulin, and the loop at the active site in triose phosphate isomerase. For instance in the introduction we referred to Keskin *et al.* who found that boundary regions of collective motions seem to act as linkages in secondary structures elements. The loops of tubulin act as these linkages, since they are dominated by low normal modes that move loops with the whole domain. Our study also confirms the finding of Gerstein *et al.* (24) that the whole protein consists of different shell regions of increasing mobility. Since most protein residues' motions are dominated by the lowest frequencies, this implies that the protein residues form clusters of rigid bodies. Another important issue is in understanding how the binding site of proteases open and close. We would like to answer the following questions: What is the mechanism for this allosteric transition? What are the roles of loops, and how do the structures of loops change during this and other transitions? Here we have made a first computational step in this direction by demonstrating that the slow motions control the loops that are most pertinent to the principle function. Our future computations will focus on the dynamical behavior of loops under certain environmental conditions and the transmission of any induced changes through the structure.

Acknowledgement

We gratefully acknowledge the financial support provided by the National Institutes of Health through grants R01GM081680, R01GM072014, and R01GM073095.

Bibliography

1. Zhang, Y., Stec, B., and Godzik, A. (2007) Between order and disorder in protein structures: analysis of "dual personality" fragments in proteins. *Structure.* **15**, 1141-1147
2. Joosten K, Cohen SX, Emsley P, Mooij W, Lamzin VS, and Perrakis A (2008) A knowledge-driven approach for crystallographic protein model completion. *Acta Crystallogr.D Biol Crystallogr* **64**, 416-424
3. Fiser A., Do RKG, and Sali A (2000) Modeling of loops in protein structures. *Prot Science* **9**, 1753-1773
4. Felts, A. K., Gallicchio, E., Chekmarev, D., Paris, K. A., Friesner, R. A., and Levy, R. M. (2008) Prediction of Protein Loop Conformations using the AGBNP Implicit Solvent Model and Torsion Angle Sampling. *J.Chem.Theory.Comput.* **4**, 855-868
5. Sellers BD, Zhu K, Zhao S, Friesner RA, and Jacobson MP (8-15-2008) Toward better refinement of comparative models: Predicting loops in inexact environments. *Proteins* **72**, 959-971
6. Olson MA, Feig M, and Brooks CL (4-15-2008) Prediction of protein loop conformations using multiscale Modeling methods with physical energy scoring functions. *J Comp Chem* **29**, 820-831
7. Peng HP and Yang AS (11-1-2007) Modeling protein loops with knowledge-based prediction of sequence-structure alignment. *Bioinformatics* **23**, 2836-2842
8. Zhu K., Pincus DL, Zhao SW, and Friesner RA (11-1-2006) Long loop prediction using the protein local optimization program. *Proteinss* **65**, 438-452
9. Xiang Z (2006) Advances in homology protein structure modeling. *Curr Protein Pept Sci* **7**, 217-227
10. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uverskyn VN, and Dunker AK (3-1-2007) Intrinsic disorder and functional proteomics

19. *Biophys J* **92**, 1439-1456

11. Panchenko AR and Madej T (2-3-2005) Structural similarity of loops in protein families: toward the understanding of protein evolution. *BMC Evol Biol* **5**, 10
12. Bos C, Lorenzen D, and Braun V (1998) Specific in vivo labeling of cell surface-exposed protein loops: Reactive cysteines in the predicted gating loop mark a ferrichrome binding site and a ligand-induced conformational change of the Escherichia coli FhuA protein. *J Bacteriol* **180**, 605-613
13. Li C, Banfield MJ, and Dennison C (1-24-2007) Engineering copper sites in proteins: Loops confer native structures and properties to chimeric cupredoxins. *J Am Chem Soc* **129**, 709-718
14. Smith JW, Tachias K, and Madison EL (12-22-1995) Protein loop grafting to construct a variant of tissue-type plasminogen activator that binds platelet integrin alpha(IIb)beta(3). *J Biol Chem* **270**, 30486-30490

15. Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu GF, van den Bedem H, Latombe JC, Halperin-Landsberg I, and Altman RB (2008) Efficient Algorithms to Explore Conformation Spaces of Flexible Protein Loops. *IEEE-ACM Trans Comp Biol Bioinformatics* **5**, 534-545

16. Krieger, Florian, Fierz, Beat, Axthelm, Fabian, Joder, Karin, Meyer, Dominique, and Kiefhaber, Thomas (2010) Intrachain diffusion in a protein loop fragment from carp parvalbumin. *Chemical Physics* **307**, 209-215

17. Li WZ, Liu ZJ, and Lai LH (1999) Protein loops on structurally similar scaffolds: Database and conformational analysis. *Biopolymers* **49**, 481-495

18. Hu, Xiaozhen, Wang, Huanchen, Ke, Hengming, and Kuhlman, Brian (11-6-2007) High-resolution design of a protein loop. *PNAS* **104**, 17668-17673

19. Keskin O, Durell SR, Bahar I, Jernigan RL, and Covell DG (2002) Relating molecular flexibility to function: A case study of tubulin. *Biophys J* **83**, 663-680

20. Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJE, and Oliva B (1-1-2004) ArchDB: automated protein loop classification as a tool for structural genomics. *Nucl Acids Res* **32**, D185-D188

21. Oliva B., Bates PA, Querol E, Aviles FX, and Sternberg MJE (3-7-1997) An automated classification of the structure of protein loops. *J Mol Biol* **266**, 814-830

22. Groban ES, Narayanan A, and Jacobson MP (2006) Conformational changes in protein loops and helices induced by post-translational phosphorylation. *Plos Comp Biol* **2**, 238-250

23. Kolodny R, Guibas L, Levitt M, and Koehl P (2005) Inverse kinematics in biology: The protein loop closure problem. *Int J Robotics Res* **24**, 151-163

24. Gerstein M and Chothia C (7-5-1991) Analysis of Protein Loop Closure - 2 Types of Hinges Produce One Motion in Lactate-Dehydrogenase. *J Mol Biol* **220**, 133-149

25. Andrec M, Snyder DA, Zhou ZY, Young J, Montellone GT, and Levy RM (11-15-2007) A large data set comparison of protein structures determined by crystallography and NMR: Statistical test for structural differences and the effect of crystal packing. *Proteins* **69**, 449-465

26. Sudarsanam S, Dubose RF, March CJ, and Srinivasan S (1995) Modeling Protein Loops Using A Phi-I+1, Psi-I Dimer Database. *Prot Science* **4**, 1412-1420

27. Street TO, Fitzkee NC, Perskie LL, and Rose GD (2007) Physical-chemical determinants of turn conformations in globular proteins. *Prot Science* **16**, 1720-1727

28. Kempf JG, Jung JY, Ragain C, Sampson NS, and Loria JP (4-20-2007) Dynamic requirements for a functional protein hinge. *J Mol Biol* **368**, 131-149

29. Bahar I, Erman B, Jernigan RL, Atilgan AR, and Covell DG (1-22-1999) Collective motions in HIV-1 reverse transcriptase: Examination of flexibility and enzyme function. *J Mol Biol* **285**, 1023-1037

30. Kurkcuoglu O, Jernigan RL, and Doruker P (1-31-2006) Loop motions of triosephosphate isomerase observed with elastic networks. *Biochemistry* **45**, 1173-1182

31. Bahar I, Atilgan AR, and Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding Des* **2**, 173-181

32. Haliloglu T, Bahar I, and Erman B (10-20-1997) Gaussian dynamics of folded proteins. *Phys Rev Lett* **79**, 3090-3093

33. Tirion MM (8-26-1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* **77**, 1905-1908

34. Flory PJ (1976) Statistical Thermodynamics of Random Networks. *Proceedings of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences* **351**, 351-380

35. James HM and Guth E (1943) Theory of the elastic properties of rubber. *J Chem Phys* **11**, 455-481

36. James HM and Guth E (1953) Statistical Thermodynamics of Rubber Elasticity. *J Chem Phys* **21**, 1039-1049

37. Kloczkowski A, Mark JE, and Erman B (1989) Chain Dimensions and Fluctuations in Random Elastomeric Networks .1. Phantom Gaussian Networks in the Undeformed State. *Macromolecules* **22**, 1423-1432

38. Skliros A, Mark JE, and Kloczkowski A (2-14-2009) Chain dimensions and fluctuations in elastomeric networks in which the junctions alternate regularly in their functionality. *J Chem Phys* **130**, 064905

39. Cui Q and Bahar I (2006) Normal Modes Analysis: Theory and applications to biological and chemical systems.

40. Jernigan RL and Kloczkowski A (2007) Packing regularities in biological structures relate to their dynamics. *Methods Mol Biol* **350**, 251-276

41. Sen TZ, Feng YP, Garcia JV, Kloczkowski A, and Jernigan RL (2006) The extent of cooperativity of protein motions observed with elastic network models is similar for atomic and coarser-grained models. *J Chem Thy Comp* **2**, 696-704

42. Atilgan AR, Durell S R, Jernigan RL, Demirel MC, Keskin O, and Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J* **80**, 505-515

43. Kim MK, Jernigan RL, and Chirikjian GS (2002) Efficient generation of feasible pathways for protein conformational transitions. *Biophys J* **83**, 1620-1630

44. Kim MK, Li W, Shapiro BA, and Chirikjian GS (2003) A comparison between elastic network interpolation and MD simulation of 16S ribosomal RNA. *J Biomol Struct Dyn* **21**, 395-405

45. Kim MK, Jernigan RL, and Chirikjian GS (2003) An elastic network model of HK97 capsid maturation. *J Struct Biol* **143**, 107-117

46. Schuyler AD and Chirikjian GS (2004) Normal mode analysis of proteins: a comparison of rigid cluster modes with C-alpha coarse graining. *J Mol Graphics Model* **22**, 183-193

47. Schuyler AD and Chirikjian GS (2005) Efficient determination of low-frequency normal modes of large protein structures by cluster-NMA. *J Mol Graphics Model* **24**, 46-58

48. Shannon CE (1949) Communication in the Presence of Noise. *Proc Inst Radio Eng* **37**, 10-21

49. Cooley JW and Tukey JW (1965) An Algorithm for Machine Calculation of Complex Fourier Series. *Math Comput* **19**, 297-301

50. Singleton RC (1969) An Algorithm for Computing Mixed Radix Fast Fourier Transform. *IEEE Trans Audio Electroacoustics* **AU17**, 93-103

51. Antoniou A (2006) Digital Signal Processing.

52. ElAli TS (2004) Discrete Systems and Digital Signal Processing with MATLAB.

53. Hayes MH (1999) Digital Signal Processing.

54. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637

55. Hubbard, S. J. and Thornton, J. M. (1993) NACCESS. *University College London*

56. Seckler, J. M., Howard, K. J., Barkley, M. D., and Wintrode, P. L. (8-18-2009) Solution structural dynamics of HIV-1 reverse transcriptase heterodimer. *Biochemistry* **48**, 7646-7655

57. Yang, L, Song G, Jernigan RL., Protein elastic network models and the ranges of cooperativity. Proc Natl Acad Sci USA 2009;106:12347-12352.